ON THE FOUNDATIONS OF PROBABILITY AND STATISTICS[1]

BY R. VON MISES

Harvard University

1. Introduction. The theory of probability and statistics which I have been upholding for more than twenty years originates in the conception that the only aim of such a theory is to give a description of certain observable phenomena, the so called mass phenomena and repetitive events, like games of chance or some specified attributes occurring in a large population. Describing means here, in the first place, to find out the relations which exist between sequences of events connected in some way, e.g. a sequence of single games and the sequence composed of sets of those games or between a sequence of direct observations and the so called inverse probability within the same field of observations. The theory is a mathematical one, like the mathematical theory of electricity, based on experience, but operating by means of mathematical processes, particularly the methods of analysis of real variables and theory of sets.

We all know very well that in colloquial language the term probability or probable is very often used in cases which have nothing to do with mass phenomena or repetitive events. But I decline positively to apply the mathematical theory to questions like this: What is the probability that Napoleon was a historical person rather than a solar myth? This question deals with an isolated fact which in no way can be considered as an element in a sequence of uniform repeated observations. We are all familiar with the fact that, e.g. the word energy is often used in every day language in a sense which does not conform to the notion of energy as adopted in mathematical physics. This does not impair the value of the precise definition of energy used in physics and on the other hand this definition is not intended to cover the entire field of daily application of the term energy.

We discard likewise the scholastic point of view displayed in a sentence of this kind: "... that both in its meaning and in the laws which it obeys, probability derives directly from intuition and is prior to objective experience." This sentence is quoted from a mathematical paper printed in a mathematical journal of 1940. The same author continues calling probability a metaphysical problem and speaking of the difficulties "which must in the nature of things always be encountered when an attempt is made to give a mathematical or physical solution to a metaphysical problem." In my opinion the calculus of probability has nothing to do with metaphysics, at any rate not more than geometry or mechanics has.

[1] Address delivered on September 11, 1940 at a meeting of the Institute of Mathematical Statistics in Hanover, N. H.

On the other hand we claim that our theory, which serves to describe observable facts, satisfies all reasonable requirements of logical consistency and is free from contradictions and obscurities of any kind. I am now going to outline the essential ideas of the theory as developed by me since 1919 and I shall have to refer as to the proof of its consistency to the recent work of A. H. Copeland, of J. Herzberg and of A. Wald. Then I will give some examples of application in order to show how the theory works and how it applies to actual problems in statistics.

2. The notion of kollektiv. The basic notion upon which the theory is established is the concept of *kollektiv*. We consider an infinite sequence of experiments or observations every one of which supplies a definite result in the form of a number (or a group of numbers in the case of a kollektiv of more than one dimension). We shall designate briefly by X the sequence of results x_1, x_2, x_3, $\cdots$. In tossing a die we get for X an endless repetition of the integers one to six, $x = 1, 2, \cdots 6$. If we are interested in death probability, we observe a large group of healthy 40 year old men and mark a one for each individual surviving his 41st aniversary and a zero for each man who dies before, so that the sequence x_1, x_2, x_3, $\cdots$ consists of zeros and ones. In a certain sense the kollektiv corresponds to what is called a *population* in practical statistics. Experience shows that in such sequences the relative frequency of the different results (one to six in the first of our examples, one and zero in the second) varies only slightly, if the number of experiments is large enough. We are therefore prompted to assume that in the kollektiv, i.e. in the theoretical model of the empirical sequences or populations, each frequency has a *limiting value*, if the number of elements increases endlessly. This limiting value of frequency is called, under certain conditions which I shall explain later, the "probability of the attribute in question within the kollektiv involved." The set of all limiting frequencies within one kollektiv is called its *distribution*.

Let me insist on the fact that in no case is a probability value attached to a single event by itself, but only to an event as much as it is the element of a well defined sequence. It happens often that one and the same fact can be considered as an element of different kollektivs. It may then be that different probability values can be ascribed to the same event. I shall give a striking example of this, which we encounter in the field of actual statistical problems, at the end of this lecture.

The objection has been made: Since all empirical sequences are obviously finite sequences, why then assume infinite kollektivs? Our answer is that any straight line we encounter in reality has finite length, but geometry is based on the notion of infinite straight lines and uses e.g. the notion of parallels which has no sense, if we restrict ourselves to segments of finite lengths. Another objection, often repeated, reads that there is a contradiction between the existence of a frequency limit and the so called Bernoulli theorem which states that sequences of any length showing a frequency say $\frac{1}{4}$ can also occur in cases for which the probability equals $\frac{1}{2}$. But it has been proved, in a rigorous way excluding any doubt, that the two statements are compatible, even by explicit construction of infinite sequences fulfilling both conditions. I would even claim that the real meaning of the Bernoulli theorem is inaccessible to any probability theory that does not start with the frequency definition of probability.

Now we are in the position to explain how our probability theory works. This sequence of zeros and ones

$$(X) \qquad 1\ 0\ 1\,|\,0\ 0\ 1\,|\,1\ 0\ 0\,|\,0\ 1\ 1\,|\,1\ 1\ 0\,|\,0\ 1\ 1\,|\,0\ 1\ 0\,|\,1\ 1\ 1\ \cdots$$

may represent the outcomes of a game of chance. The ones show gains, the zeros losses for one of the two players. If we separate the terms of X into groups of three digits and replace each group by a single one or zero according to the majority of terms within the group, we get a new sequence

$$(X') \qquad 1\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ \cdots$$

which represents the gains and losses in sets of three games. Our task is now to compute the distribution, i.e. the limiting frequencies of zeros and ones in this new sequence X', assuming the two frequencies in X are known. A sequence can formally be considered as a unique number like a decimal fraction with an infinite number of digits. Then the transition from X to X' can be called a *transformation of a number* $X' = T(X)$. As our sequences have to fulfill certain conditions Copeland calls the sequences X, X' admissible numbers. What I just quoted was of course a very special example of a transformation of a number. But we have to emphasize that all problems dealt with in probability theory, without any exception, have this unique form: The distribution or the limiting frequencies in certain sequences are given, other sequences are derived from the given ones by certain operations, and the distributions in these derived sequences have to be computed. In other words: *Probability theory is the study of transformations of admissible numbers, particularly the study of the change of distributions implied by such transformations.*

We know four and only four simple, i.e. irreducible transformations or *four fundamental operations*. They are called selection, mixing, partitioning and combination. By combining these basic processes we can settle all problems in probability theory. The formal, mathematical difficulties in carrying out the computation of the new distributions may become very serious in certain cases, particularly if we have to apply an infinite number of transformations (asymptotic problems). But, in the clearly defined framework of this theory no space is left for any metaphysical speculations, for ideas about sufficient reason or insufficient reason, for notions like degree of evidence or for a special kind of probability logic and so on. And further no modification is needed for handling usual statistical problems: Terms like inverse probability, likelihood, confidence degrees, etc. are justified and admitted only as far as they are capable of being reduced to the basic notion of kollektiv and distribution within a kollektiv. I will give some more details to this point later. Meanwhile let me turn to a general question which, in a certain way, is the crucial point in establishing the new probability theory.

3. Place selections and randomness. It is obvious that we have to restrict still further the notion of kollektiv or the field of sequences which can be considered as the objects of a probability investigation. The successive outcomes of a game of chance differ very clearly from any regular sequence as defined by a simple arithmetical law, e.g. the regularly alternating sequence 0 1 0 1 0 1 0 1 $\cdots$. A typical property which singles out the irregular or random sequences and which has to be reproduced in every probability theory is that, if p is the probability of encountering a one in the sequence, then p^2 is the probability of two ones following each other immediately. Any probability theory has to introduce an axiom which enables us to deduce this theorem and others of a similar type. The question is only how to find a sufficiently general and consistent form for it. The procedure I have chosen consists in using a special kind of transformation of a sequence, which I call a *place selection*.

A place selection is defined by an infinite set of functions $s_n(x_1, x_2, \cdots x_{n-1})$ where $x_1, x_2, x_3, \cdots$ are the digits of an admissible number or a kollektiv and s_n has one of the two values zero or one. Here $s_n = 1$ means that the nth digit of the sequence is retained, $s_n = 0$ means that it is discarded. The decision about retaining or discarding the nth elements depends as you see, only on the preceding values $x_1, x_2, \cdots x_{n-1}$, but not on x_n or the following digits. Example of a place selection:

$$s_n = 1, \text{ if } x_{n-1} = 0 \text{ for prime numbers } n,$$

$$\text{if } x_{n-1} = 1 \text{ for } n \text{ not prime},$$

$$s_1 = 1, \text{ and } s_n = 0 \text{ in all other cases.}$$

Experience shows that, if we apply such a place selection to the sequence X of outcomes of a game of chance, we get a new, selected sequence $S(X)$ in which the frequencies of gains and losses are about the same as in X. This fact or the practical *impossibility of a gambling system* suggests the adoption of the following procedure in handling transformations of admissible numbers.

First, if within a certain investigation the transformation applied to X is a place selection, we assume that the distribution in $X' = S(X)$ is the same as in X: distr $S(X) =$ distr X. Second, if a general transformation T is applied to X, say $X' = T(X)$, then we examine whether the existence of a place selection S that changes the distribution in X' (so as to have distr $S(X') \neq$ distr X') implies the existence of a place selection S_1 that would affect the distribution in X (so as to give distr $S_1(X) =$ distr X). If this is the case, we say that X' is a kollektiv, provided that the original sequence X was considered to be a kollektiv. Take e.g. for X the sequence resulting from tossing a die endlessly, and call $p_1, p_2, \cdots p_6$ the limiting frequencies of the six possible outcomes 1, 2, $\cdots$ 6. The transformation T may consist in replacing every 1 in the sequence X by a

2, every 3 by a 4, and every 5 by a 6. The new sequence consists of only three different kinds of elements 2, 4, 6 and therefore its distribution includes only three values p'_2, p'_4, p'_6 where evidently $p'_2 = p_1 + p_2$ etc. Here it is almost obvious that if a place selection applied to X' changes the value of p'_2, the same selection if applied to X must change either p_1 or p_2. So, if the original sequence X was considered as a kollektiv, X' has to be admitted too.

Now the question arises whether this procedure is in itself consistent or whether it can lead to contradictions. We were concerned up to now with kollektivs the elements of which belong to a finite set of distinct numbers $e_1, e_2, \cdots e_k$ and the distributions of which are therefore defined by k non-negative values $p_1, p_2, \cdots p_k$ with the sum 1. In this case it was pointed out by Wald and by Copeland that, if an arbitrary distribution and an arbitrary countable set Σ of place selections are given, there exists a continuum of sequences every one of which has the given distribution, which is not affected by any place selection belonging to Σ. Now it may be supposed that in a concrete problem a sequence X' is derived from a sequence X by a finite number of fundamental operations involving a finite set Σ' of place selections. Another finite set Σ'' may consist of selections employed in establishing that certain sequences used in the derivation of X' are "combinable" ones. Finally an arbitrary countable set Σ of selections S may be assumed. According to our procedure we have shown that to any place selection S which affects the distribution in X' corresponds a certain S_1 which, when applied to X, changes the distribution of X. All these S_1 corresponding to the elements S of Σ form a countable set Σ_1. Now the set Σ_2 including Σ', Σ'', Σ_1 and also including all products of two of its own elements is a countable set too. What we use in computing the distribution of X' is only the fact that the given sequence X is unaffected by the selections that are elements of Σ_2. It follows from the above quoted results that we can substitute for X a numerically specified sequence and carry out all operations upon this specified sequence. So it is proved that no contradiction can arise in computing the final probability according to our conception.

I cannot enter here into a discussion of the more complicated case where the range within which the elements of a kollektiv vary, is an infinite one, either a countable set or a continuum. All principal problems connected with establishing the notion of kollektiv can be settled satisfactorily, at any rate, by considering those general forms of sequences as limiting cases of kollektivs with a finite set of attributes.

4. Example: Set-of-games problem. I want to present now a simple, but instructive example to show how the theory works and what task a mathematical foundation of the calculus of probability has to achieve. Let us recall the two sequences X and X' composed of zeros and ones of which we spoke above. The first represented the outcomes of a sequence of single games, the second the outcomes of triple sets of those games. If X is considered as a kollektiv with given probabilities p and q for one and zero, it is easy to deduce the corresponding values p' and q' for X' and to show that X' is a kollektiv too. We begin by carrying out three selections which single out from the original sequence x_1, x_2, $x_3 \cdots$ first, the elements x_1, x_4, x_7, $\cdots$ second, the elements x_2, x_5, x_8, $\cdots$ and third, the elements x_3, x_6, x_9, $\cdots$. It can be shown by means of certain further place selections that these three kollektivs which we call X_1, X_2, X_3 are combinable. That means that combining the corresponding elements of the three sequences like $x_1x_2x_3$, $x_4x_5x_6$, $x_7x_8x_9$, $\cdots$ leads to a new three dimensional kollektiv X_0 in which each permutation of three digits 0 and 1, has a probability equal to the corresponding product of p- and q-factors. For instance the probability of encountering the group 111 is p^3 and for the group 110 it is p^2q. Now we operate a mixing upon X_0 by collecting all permutations with two or three ones. We find in a well known way the sum $p^3 + 3p^2q$ for the probability p' of ones in the sequence X'. So far the result is very well known and can be reached—in my opinion, in a very incomplete and unsatisfactory way—also by the classical methods.

But what I want to discuss here is a slightly modified question. If the sequence X means gains and losses for single games and if the arrangement for sets of three games is made as indicated before, then in a real play the gains and losses of sets are counted in a different way. For, if the first two games of a set are both won or lost by the same player, the fate of the set is decided and there is no sense to play the third game. So the loss of the second set in our example will already be recognized after the fifth game and the actual sixth game will be considered as the first game of the third set. In this way the original sequence X decomposed into groups of two or three games

(X) 1 0 1 | 0 0 | 1 1 | 0 0 | 0 1 1 | 1 1 | 0 0 | 1 1 | 0 1 0 | 1 1 | $\cdots$

leads to a new sequence X''

(X'') 1 0 1 0 1 1 0 1 0 1 $\cdots$

which is obviously different from X'. Everyone familiar with the usual handling of the probability concept will say that in X'' the probabilities of zeros and ones must be the same as in X'. But a mathematical foundation of theory of probability, if it deserves this name, has to clear up the question: From what principles or particular assumptions and by what inferences may we deduce the equality of the limiting frequencies in X' and X''?

There is no difficulty in solving this problem from the point of view of the frequency theory. We have only to apply somewhat different place selections instead of the above used which lead to the kollektivs X_1, X_2, X_3. I showed elsewhere how the general set-of-games problem can be satisfactorily treated in this way. Here I want to stress only that the problem as a whole is completely inaccessible by any of the other known approaches to probability theory. The classical point of view which starts with the notion of equally likely cases and rests upon a rather vague idea of the relationship between probability and sequences of events does not even allow the formulation of the problem. In the so called modernized classical theory, as proposed by Fréchet, probabilities are defined as "physical magnitudes of which frequencies are measures." Fréchet would say that the frequencies both in X' and in X'' are measures of the same quantity. But why? We face here obviously a mathematical question which cannot be settled by referring to physical facts. It is clear that the equality of the distributions in the two sequences X' and X'' is due to the randomness or irregularity of the original sequence X. No theory which does not take in account the randomness, which avoids referring to this essential property of the sequences dealt with in probability problems, can contribute anything toward the solution of our question.

I have to make some special remarks about the so-called measure theory of probability.[2]

5. Probability as measure. Up to now we have been concerned only with the simplest type of kollektivs, namely, with those sequences the elements of which belong to a finite set of numbers so as to have a distribution consisting of a finite number of finite probabilities with the sum 1. It may be true that all practical problems, in a certain sense, fall into this range. For, the single result of an observation is always an integer, the number of smallest units accessible to the actual method of measuring. Nevertheless in many cases it is much more useful to adopt the point of view that the possible outcomes of an experiment belong to a more general set of numbers, e.g. to a continuous segment or any infinite variety. If we include the case of kollektivs of more than one dimension, we have to consider a point set in a k-dimensional space (where even k may be infinite) as the label set or attribute set of the kollektiv. In order to define the probability in this case we have to choose a subset A of the label set and to count among the first n elements the number n_A of those elements the attributes of which fall into A. Then the quotient $n_A : n$ is the frequency, and its limiting value for n infinite will be called the probability of the attribute falling into A within the given kollektiv.

It was rightly stressed by many authors that in the case of an infinite label set some additional restrictions must be introduced. In particular A. Kolmogoroff set up a complete system of such restrictions. We cannot ask for the existence of the limiting frequency in any arbitrary subset A. It will be sufficient to assume that the limit exists for a certain Körper or a certain additive family of subsets. If it exists for two mutually exclusive subsets A and B, the limit corresponding to $A + B$ will be, by virtue of the original definition, the sum of the limits connected with A and B. We can now insert a further axiom involving the complete additivity of the limiting values. So we arrive at the statement

[2] What I call measure theory here is essentially that proposed by Kolmogoroff in his pamphlet of 1933. As to the new theory developed by Doob in his following paper (where instead of the label space the space of all logically possible sequences is used in establishing the measures) see my comment on page 215.

that probability is the measure of a set. All axioms of Kolmogoroff can be accepted within the framework of our theory as a part of it, but in no way as a substitute for the foregoing definition of probability.

Occasionally the expression probability as measure theory is used in a different sense. One tries to base the whole theory on the special notion of a set of measure zero. One of the basic assumptions in my theory is that in the sequence of results we obtain in tossing a so called correct die the frequency, say of the point 6, has a certain limiting value which equals 1/6. A different conception consists in stating that anything can happen in the long run with a correct die, even that an uninterrupted sequence of six's or an alternating sequence of two's and four's or so on may appear. Only all these events which do not lead to the limiting frequency 1/6 form, together as a whole, a set of events of measure zero. Instead of my assumption: the limiting value is 1/6 we should have to state: It is almost certain that a limit exists and equals 1/6. Nothing can be said against such an alluring assumption from an empirical standpoint, since actual experience extends in no case to an infinite range of observations. The only question is whether the asumption is compatible with a complete and consistent theory. I cannot see how this may be achieved. Before saying that a set has measure zero we have to introduce a measure system which can be done in innumerable ways. If e.g. we denote the outcome six by a one and all other outcomes 1 to 5 by zero, we get as the result of the game with a die an infinite sequence of zeros and ones. It has been shown by Borel that according to a common measure system the set of all 0, 1 sequences which do not have the limiting frequency $\frac{1}{2}$ has the measure zero. In this way it turns out to be almost certain that the limiting frequency of the outcome six in the case of a correct die is $\frac{1}{2}$. Other values for the limit can be obtained by a similar inference. It is a correct but misleading idea that the measure zero is unaffected by a regular (continuous) transformation of the assumed measure system, since in our field of problems different measures which are not obtained from one another by a regular transformation have equal rights. So, saying that a certain set has the measure zero makes in our case no more sense than to state that an unknown length equals 3 without indicating the employed unit.

In recapitulating this paragraph I may say: First, the axioms of Kolmogoroff are concerned with the distribution function within one kollektiv and are *supplementary to my theory, not a substitute for it.* Second, using the notion of measure zero in an absolute way without reference to the arbitrarily assumed measure system, *leads to essential inconsistencies.*

6. Statistical estimation. Let me now turn to the last point, the application of probability theory to one of the most widely discussed questions in today's statistical research: the so-called estimation problem. Many strongly divergent opinions are facing each other here. I think that the probability theory based on the notion of kollektiv is best able to settle the dispute and to clear up the difficulties which arose in the controversies of different writers.

We may, without loss of generality, restrict ourselves to the simplest case of a single statistical variable x and a single parameter ϑ, where x of course may be the arithmetical mean of n observed values. Here (and likewise in the case of more variables and more parameters) we have to distinguish carefully among four different kollektivs which are simultaneously involved in the problem. The range within which both x and ϑ vary will be assumed to be a continuous interval so that all distributions will be given by probability densities.

The first kollektiv we deal with is a one-dimensional one where the probability of x falling into the interval x, $x + dx$ depends on x and on a parameter ϑ. If

$$p(x \mid \vartheta) \tag{1}$$

denotes the corresponding density and the limits A, B within which x possibly falls depend on ϑ too, we have

$$\int_{A(\vartheta)}^{B(\vartheta)} p(x \mid \vartheta)\, dx = 1 \qquad \text{for each } \vartheta. \tag{1'}$$

In order to fix the ideas we may imagine that the first kollektiv consists in drawing a number x out of an urn and that ϑ characterizes the contents of the urn. Asking for an estimate of ϑ implies the assumption that different possible urns are at our reach every one of which can be used for drawing the x. The ϑ values for the different urns fall into a certain interval C, D. It is usual to suppose that the urns are picked out at random so as to give another one-dimensional kollektiv with the independent variable ϑ. Let $p_0(\vartheta)\, d\vartheta$ be the probability of picking an urn with the characteristic value falling into the interval ϑ, $\vartheta + d\vartheta$. This density

$$p_0(\vartheta) \tag{2}$$

is often called the *prior* or *a priori* probability of ϑ. As the range within which ϑ varies is confined by the constants C and D, we have obviously

$$\int_C^D p_0(\vartheta)\, d\vartheta = 1. \tag{2'}$$

Now from these two one-dimensional kollektivs with the variables x in the first, ϑ in the second, we deduce by combination (multiplication) a two-dimensional kollektiv with the density function

$$P(\vartheta, x) = p_0(\vartheta) \cdot p(x \mid \vartheta). \tag{3}$$

The individual experiment which forms the element of this third kollektiv consists of picking at random an urn and drawing afterwards from this urn. Both x and ϑ are now independent variables (attributes of the kollektiv) and it is easy to see that it follows from (1) and (2)

$$\int_C^D \int_{A(\vartheta)}^{B(\vartheta)} P(\vartheta, x)\, dx\, d\vartheta = \int_C^D p_0(\vartheta)\, d\vartheta \int_{A(\vartheta)}^{B(\vartheta)} p(x \mid \vartheta)\, dx = 1. \tag{3'}$$

We will return later to this two-dimensional kollektiv. Let us, first, derive from it, by applying the operation of partitioning (Teilung), our fourth and last kollektiv which is one-dimensional again. Partitioning means that we drop from the sequence of experiments which form the third kollektiv all those for which the x-value falls outside a certain interval x, $x + dx$; and that in this way we consider a partial sequence of experiments with only the one variable ϑ. The distribution of ϑ-values within this sequence with quasi-constant x is given, according to the well known rule of division or rule of Bayes (a rule which can be proved mathematically) by[3]

$$p_1(\vartheta \mid x) = \frac{P(\vartheta, x)}{\int_C^D P(\vartheta, x)\, d\vartheta} = c(x)\, p_0(\vartheta)\, p(x \mid \vartheta). \tag{4}$$

It follows immediately that

$$\int_C^D p_1(\vartheta \mid x)\, d\vartheta = 1. \tag{4'}$$

This function p_1 of ϑ depending on the parameter x is generally called the *posterior* or *a posteriori* probability of ϑ.

If $p_1(\vartheta \mid x)$ can be computed according to the formula (4), every question concerning the "presumable" value of ϑ as drawn from the outcome x of an experiment is completely answered. We can find indeed, by integration the probability which corresponds to any part of the interval C, D of ϑ and so the estimation problem is definitely solved. But the trouble is that in most cases of practical application nothing or almost nothing is known about the prior probability $p_0(\vartheta)$ which appears as a factor in the expression of p_1. Hence arises the new question: *What can we say about the ϑ-values without having any information about its prior probability?* This is the estimation problem as it is generally conceived today.

The first successful approach to the answering of this question was made by Gauss. If we do not know p_1, we know however, except for a constant factor, the quotient p_1/p_0, posterior probability to prior probability which equals $cp(x \mid \vartheta)$. The maximum of this quotient must be greater than one, since the average values of both p_0 and p_1 are the same. So the maximum means the point of the greatest increase produced by the observed experimental value of x upon the probability of ϑ. It seems reasonable to assume the ϑ-value for which the ratio p_1/p_0 reaches its maximum as an estimate for ϑ: It is the value upon which the greatest emphasis is conferred by the observation. This idea, originally proposed by Gauss in his theory of errors, has been later developed chiefly by R. A. Fisher, and is known today as the maximum likelihood method. Calling the ratio p_1/p_0 likelihood seems indeed an adequate nomenclature.

[3] For brevity Bayes' rule is employed in the text as in the case of a discontinuous distribution. The correct procedure in the case of a continuous x would require that we first use finite intervals and then pass to the limit.

The method of estimation used most frequently today is not the maximum likelihood method, but the so called confidence interval method, inaugurated by R. A. Fisher and now successfully extended and applied by J. Neyman. This method uses the third of the above mentioned kollektivs instead of the fourth, i.e. the two-dimensional probability $P(\vartheta, x)$. At first sight it seems hopeless to use this function which includes the unknown prior probability $p_0(\vartheta)$ as a factor. But it turns out as Neyman has shown[4] (and this is the decisive idea of the confidence interval method) that we can indicate in the x, ϑ-plane special regions for which the probability $\int\int P(\vartheta, x)\, dx\, d\vartheta$ is independent of $p_0(\vartheta)$. In fact, if we point out for every ϑ such an interval x_1, x_2 as to have

$$\int_{x_1(\vartheta)}^{x_2(\vartheta)} p(x \mid \vartheta)\, dx = \alpha, \qquad 0 < \alpha < 1, \tag{5}$$

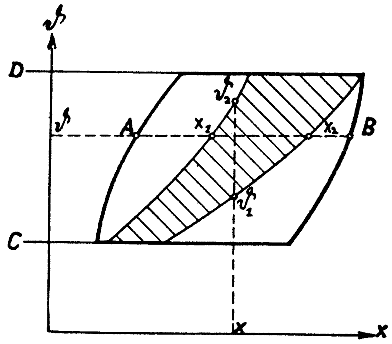

FIG. 1

it follows immediately from (2) and (5) for the region covered by these intervals

$$\int_C^D \int_{x_1(\vartheta)}^{x_2(\vartheta)} P(\vartheta, x)\, dx\, d\vartheta = \int_C^D p_0(\vartheta)\, d\vartheta \int_{x_1(\vartheta)}^{x_2(\vartheta)} p(x \mid \vartheta)\, dx = \alpha. \tag{6}$$

For given α the intervals can be chosen in different ways. If we choose $x_1 = A$ for $\vartheta = C$ and $x_2 = B$ for $\vartheta = D$, we get a strip or belt, as shown in Fig. 1 which supplies for every given x a smallest value ϑ_1 and a greatest value ϑ_2. The definition of our third kollektiv leads to the conclusion: *If we predict each time a certain x is observed that ϑ lies between the corresponding ϑ_1 and ϑ_2, then the probability is α that we are right, whatever the prior probability may be.*[5] It is

[4] J. Neyman, *Roy. Stat. Soc. Jour.*, Vol. 97 (1934), pp. 590–92.

[5] After my lecture Dr. A. Wald called my attention to Neyman's suggestion; namely that this statement can be generalized by admitting that the infinite sequence of ϑ-values which results from picking out successively the urns for drawing a number x, does not fulfill the conditions of a kollektiv. So, instead of the terms "whatever the prior probability may be" we can say "whatever the method of picking out the urns may be." In fact, let us consider the case where ϑ can assume only a finite number of values ϑ_1, ϑ_2, $\cdots$ ϑ_k. Among the n first trials let n_κ be the number of cases where $\vartheta = \vartheta_\kappa$ and $n'_\kappa \leqq n_\kappa$ the number of cases where $\vartheta = \vartheta_\kappa$ and x falls into the interval $x_1(\vartheta_\kappa)$, $x_2(\vartheta_\kappa)$. The relative

understood that in this argument both x and ϑ are variables the values of which may change from one trial to the next. I cannot agree with the statement, which is often made, that x only is a variable and ϑ a constant or that we are only interested in one specified value of ϑ. In no way is it possible, in the framework of the confidence limits method, to avoid the idea of a so-called superpopulation, i.e. the existence of a manifold of urns every one of which forms a kollektiv.[6] Thus no contradiction and no antagonism exists between this method and the Bayes formula. Only a different kollektiv, a two-dimensional instead of a one-dimensional, is here considered.

I have no time to enter here in a discussion of the very interesting developments of Neyman's theory which are intended to supply additional conditions in order to determine the arbitrary choice of the x-intervals in a unique way. May I only mention that what is called in Neyman's theory the probability of a second type error in testing the hypothesis $\vartheta = \vartheta_0$ is given by the expression

$$\int_C^D \int_{x_1(\vartheta_0)}^{x_2(\vartheta_0)} P(\vartheta, x)\, dx\, d\vartheta = \int_C^D p_0(\vartheta)\, d\vartheta \int_{x_1(\vartheta_0)}^{x_2(\vartheta_0)} p(x \mid \vartheta)\, dx. \tag{7}$$

If we want to determine the confidence belt or the intervals x_1, x_2 in such a way as to minimize this expression independently of the function $p_0(\vartheta)$, we obtain Neyman's maximum power condition

$$\int_{x_1(\vartheta_0)}^{x_2(\vartheta_0)} p(x \mid \vartheta)\, dx \equiv F(\vartheta, \vartheta_0) = \text{min. for each pair } \vartheta, \vartheta_0. \tag{8}$$

This condition, it is well known, cannot be fulfilled under general assumptions for $p(x \mid \vartheta)$. Moreover the above-mentioned boundary conditions $x_1(C) = A(C)$ and $x_2(D) = B(D)$ (or similar ones in other cases) have to be considered too. If they are not satisfied, the statement which can be made with probability α would include the prediction that certain x-values are impossible. Except for this case the above formulated theorem is equally valid for every region determined according to (5).

It is clear that if the original distribution is given by a regular, slightly varying function $p(x \mid \vartheta)$, the confidence limits method cannot give very substantial results. Let us take e.g. for $p(x \mid \vartheta)$ the uniform distribution

$$p(x \mid \vartheta) = 1/\vartheta \text{ for } 0 \leqq x \leqq \vartheta, \qquad 0 \leqq \vartheta \leqq 1. \tag{9}$$

frequency of correct predictions is then $(n_1' + n_2' + \cdots n_k')$: n where n equals $n_1 + n_2 + \cdots n_k$. If n tends to infinity, at least one part of the n_κ must become infinite. For those the limit of $n_\kappa' : n_\kappa$ tends to α according (5) while the other terms (with finite n_κ and n_κ') have no influence. So the limiting value of the frequency $(n_1' + n_2' + \cdots n_k')$: n equals in any event α. This generalization does not apply, if we ask for the probability of a second type error of the hypothesis $\vartheta = \vartheta_0$. Here the existence of the prior probability p_0 is essential.

[6] According to the generalization supplied by Neyman's point of view (*Phil. Trans. Roy. Soc.*, Vol. A-236 (1937), pp. 333–380) which is discussed in footnote 5, the superpopulation does not necessarily satisfy the conditions of a kollektiv.

We have here $A = 0$, $B = \vartheta$, $C = 0$, $D = 1$ and the domain in which x and ϑ vary is the 45° right triangle shown in Fig. 2. Whatever $p_0(\vartheta)$ may be, the integral of $p(\vartheta, x) = p_0(\vartheta) \cdot p(x \mid \vartheta)$ over this domain is 1 and if we omit the part of the triangle on the left of the straight line $x = (1 - \alpha)\vartheta$, the integral over the remaining part is α. For $\alpha = 0.90$, a statement which can be made with a probability of 90% reads: The value of ϑ lies between x and $10x$. On the other hand we know from the very beginning with 100% certainty that ϑ lies between x and 1, so that for $x \geqq 0.1$ the statement is futile. (If one chooses as confidence belt the part on the left of the straight line $x = \alpha\vartheta$, the statement would run: ϑ lies between 1.1 x and 1 and values of x greater than 0.9 are impossible.) If we apply in this case the Bayes formula, we find that the outcome depends to the highest extent on what is known about the prior probability $p_0(\vartheta)$.

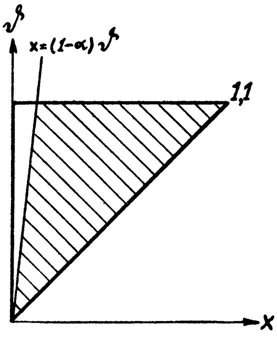

FIG. 2

In most cases however which present themselves in practical statistics the original density function $p(x \mid \vartheta)$ has a different character from that assumed in (9). It depends generally on an integer n and the distribution is concentrated more and more when n increases. (We may define here concentration as standard deviation tending towards zero. The integer n means in general the number of basic experiments). We have e.g. in the so-called Bayes problem where x is the arithmetical mean of n observations the asymptotic expression for p:

$$p(x \mid \vartheta) \sim \sqrt{\frac{n}{2\pi\vartheta(1 - \vartheta)}}\, e^{-\frac{1}{2}n(x-\vartheta)^2/\vartheta(1-\vartheta)} \tag{10}$$

$$0 \leqq \vartheta \leqq 1, \qquad 0 \leqq x \leqq 1.$$

If we denote by Φ the probability integral

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2}\, du, \tag{11}$$

the x-intervals corresponding to a given probability value α are defined by

$$x_1 = \vartheta - \xi, \qquad x_2 = \vartheta + \xi \qquad \text{where } \Phi\left(\xi \sqrt{\frac{n}{2\vartheta(1 - \vartheta)}}\right) = \alpha. \tag{12}$$

If n has a large value, the ξ's are very small and we get a narrow belt along the straight line $x = \vartheta$ as shown in Fig. 3 for $\alpha = 0.90$ and n about 100. The prediction which can be made with the probability α reads approximately

$$(13) \qquad x - \eta \leqq \vartheta \leqq x + \eta \qquad \text{where } \Phi\left(\eta \sqrt{\frac{n}{2x(1-x)}}\right) = \alpha.$$

On the other hand it is well known that in this case the Bayes formula supplies a posterior probability $p_1(\vartheta \mid x)$ which turns out to be more and more independent of the prior probability $p_0(\vartheta)$ when n increases. It has been shown that the asymptotic expression for $p_1(\vartheta \mid x)$ whatever $p_0(\vartheta)$ may be, is

$$(14) \qquad p_1(\vartheta \mid x) \sim \sqrt{\frac{n}{2\pi x(1-x)}}\, e^{-\frac{1}{2}n(\vartheta - x)^2/x(1-x)}.$$

It follows that, on the basis of the Bayes formula, we can predict for every single value of x with the probability α that ϑ lies between the above given limits (13). This is more than the confidence limits method supplies, but the result is subjected to the restriction that $p_0(\vartheta)$ is a continuous function. However, for large values of n (generally this means for large numbers of basic experiments) the outcomes of both methods are essentially the same.

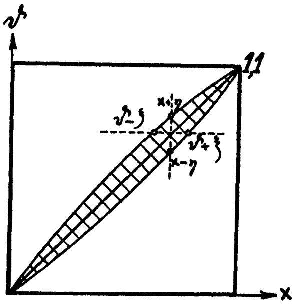

FIG. 3

Let me recapitulate in three brief sentences the essential results we have found in the problem of estimation.

1. There is no contradiction of any kind between the Bayes formula and the confidence limits method and no difference at all in the underlying probability concept. In both methods the idea of a sort of "super-population" is used. Only two different kollektivs are considered in both cases.

2. If the original distribution has a regular, slightly varying density function $p(x \mid \vartheta)$, the Bayes method gives a complete answer when the prior probability is known and no answer when it is unknown. The confidence limits method gives in both cases a definite solution; it lies in the nature of things that the solution cannot be very substantial if $p(x, \vartheta)$ is only slightly varying.

3. If the original distribution $p(x \mid \vartheta)$ depends on a further parameter n and becomes concentrated more and more with increasing n, both approaches give, for large n, asymptotically about the same results.

It is not intended by these remarks to impair the value of the confidence limits method which both from theoretical and from practical point of view deserves our attention. But the rather inconceivably aggressive attitude towards the Bayes' theory as displayed by a number of statisticians, which, however, does not include J. Neyman, turns out to be completely unfounded.