THEORY PROBAB. APPL.
Vol. 48, No. 2, pp. 191–220

FROM THE HERITAGE OF A. N. KOLMOGOROV: THE THEORY OF PROBABILITY*

A. N. KOLMOGOROV

DOI. 10.1137/S0040585X97980336

In 1956 the AN USSR Publishing House published three volumes of the monograph *Mathematics*: *Its Content*, *Methods*, *and Meaning* which was elaborated by the Steklov Mathematical Institute RAN. A. D. Aleksandrov, A. N. Kolmogorov, and M. A. Lavrent'ev were the members of the editorial board. In order for the mathematical community to have an opportunity to discuss the monograph, 350 copies of it were printed in 1953 as a manuscript.

Kolmogorov's idea was that it would be good to have two books: a first which informally was planned as "Anti-Courant" (see the introduction to the 3rd Russian edition of R. Courant and H. Robbins, *What is Mathematics*? *An Elementary Approach to Ideas and Methods*, Oxford University Press, London, New York, 1996), i.e., a book for everybody who wants in vivid and simple form to get to know the elements of higher mathematics, to test the level of his abilities in mathematics, and, for a young reader, to consider choosing mathematics as his profession, and a second book "intended for more advanced readers including ourselves, mathematicians, who very often are helpless in estimating future trends of their science as a whole." Finally, three volumes containing 20 chapters showed the best correlation with the first of the variants indicated above. This follows additionally from the introduction which says that "the purpose of the author was to acquaint a wide Soviet circle with the content and methods of separate mathematical disciplines, their material resources, and paths of development."

In Chapter XI of the second volume of this monograph, the Kolmogorov paper was published, which is reprinted in the present jubilee issue together with the Khinchin referee report and selected correspondence of A. D. Aleksandrov (Editor-in-Chief of the monograph) with A. N. Kolmogorov, which are interesting both for their view on the content of the variant of the paper presented by Kolmogorov and for the philosophical and methodological aspects of probability theory.

A. N. Shiryaev

THE THEORY OF PROBABILITY

1. The laws of probability. The simplest laws of natural science are those that state the conditions under which some event of interest to us will either certainly occur or certainly not occur; i.e., these conditions may be expressed in one of the following two forms:

1. If a complex (i.e., a set or collection) of conditions S is realized, then event A certainly occurs;

2. If a complex of conditions S is realized, then event A cannot occur.

In the first case the event A, with respect to the complex of conditions S, is called a "certain" or "necessary" event, and in the second an "impossible" event. For example, under atmospheric pressure and at temperature t between 0° and 100° (the complex of conditions S) water necessarily occurs in the liquid state (the event A_1 is certain) and cannot occur in a gaseous or solid state (events A_2 and A_3 are impossible).

*Reprinted from A. D. Aleksandrov, A. N. Kolmogorov, and M. A. Lavrent'ev, *Mathematics*: *Its Content*, *Methods*, *and Meaning, 2nd ed., Vol.* 2, MIT Press, Cambridge, MA, 1969, Part 4, Chap. XI, pp. 229–264, with permission of the MIT Press.

http://www.siam.org/journals/tvp/48-2/98033.html

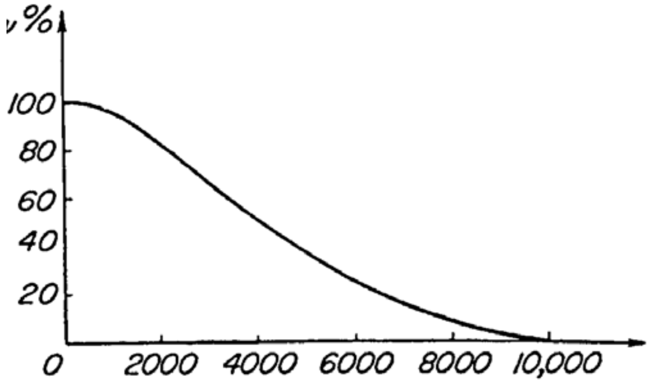

FIG. 1.

An event A, which under a complex of conditions S sometimes occurs and sometimes does not occur, is called random with respect to the complex of conditions. This raises the question: Does the randomness of the event A demonstrate the absence of any law connecting the complex of conditions S and the event A? For example, let it be established that lamps of a specific type, manufactured in a certain factory (condition S) sometimes continue to burn more than 2,000 hours (event A), but sometimes burn out and become useless before the expiration of that time. May it not still be possible that the results of experiments to see whether a given lamp will or will not burn for 2,000 hours will serve to evaluate the production of the factory? Or should we restrict ourselves to indicating only the period (say 500 hours) for which in practice all lamps work without fail, and the period (say 10,000 hours) after which in practice all lamps do not work? It is clear that to describe the working life of a lamp by an inequality of the form $500 \leqq T \leqq 10,000$ is of little help to the consumer. He will receive much more valuable information if we tell him that in approximately 80% of the cases the lamps work for no less than 2,000 hours. A still more complete evaluation of the quality of the lamps will consist of showing for any T the percent $\nu(T)$ of the lamps which work for no less than T hours, say in the form of the graph in Figure 1.

The curve $\nu(T)$ is found in practice by testing with a sufficiently large sample (100–200) of the lamps. Of course, the curve found in such a manner is of real value only in those where it truly represents an actual law governing not only the given sample but all the lamps manufactured with a given quality of material and under given technological conditions; that is, only if the same experiments conducted with another sample will give approximately the same results (i.e., the new curve $\nu(T)$ will differ little from the curve derived from the first sample). In other words, the statistical law expressed by the curves $\nu(T)$ for the various samples is only a reflection of the law of probability connecting the useful life of a lamp with the materials and the technological conditions of its manufacture.

This law of probability is given by a function $\mathbf{P}(T)$, where $\mathbf{P}(T)$ is the probability that a single lamp (made under the given conditions) will burn no less than T hours.

The assertion that the event A occurs under conditions S with a definite probability

$$\mathbf{P}(A/S) = p$$

amounts to saying that in a sufficiently long series of tests (i.e., realizations of the complex of conditions S) the frequencies

$$\nu_r = \frac{\mu_r}{n_r}$$

of the occurrence of the event A (where n_r is the number of tests in the rth series, and μ_r is the number of tests of this series for which event A occurs) will be approximately identical with one another and will be close to p.

The assumption of the existence of a constant $p = \mathbf{P}(A/S)$ (objectively determined by the connection between the complex of conditions S and the event A) such that the frequencies ν get closer "generally speaking" to p as the number of tests increases, is well borne out in practice for a wide class of events. Events of this kind are usually called *random* or *stochastic*.

This example belongs to the laws of probability for mass production. The reality of such laws cannot be doubted, and they form the basis of important practical applications in statistical quality control. Of a similar kind are the laws of probability for the scattering of missiles, which are basic in the theory of gunfire. Since this is historically one of the earliest applications of the theory of probability to technical problems, we will return below to some simple problems in the theory of gunfire.

What was said about the "closeness" of the frequency ν to the probability p for a large number n of tests is somewhat vague; we said nothing about how small the difference $\nu - p$ may be for any n. The degree of closeness of ν to p is estimated in section 3. It is interesting to note that a certain indefiniteness in this question is quite unavoidable. The very statement itself that ν and p are close to each other has only a probabilistic character, as becomes clear if we try to make the whole situation precise.

2. The axioms and basic formulas of the elementary theory of probability. Since it cannot be doubted that statistical laws are of great importance, we turn to the question of methods of studying them. First of all one thinks of the possibility of proceeding in a purely empirical way. Since a law of probability exhibits itself only in mass processes, it is natural to imagine that in order to discover the law we must conduct a mass experiment.

Such an idea, however, is only partly right. As soon as we have established certain laws of probability by experiment, we may proceed to deduce from them new laws of probability by logical means or by computation, under certain general assumptions. Before showing how this is done, we must enumerate certain basic definitions and formulas of the theory of probability.

From the representation of probability as the standard value of the frequency $\nu = m/n$, where $0 \leqq m \leqq n$, and thus $0 \leqq v \leqq 1$, it follows that the probability $\mathbf{P}(A)$ of any event A must be assumed to lie between zero and one[1]

$$0 \leqq \mathbf{P}(A) \leqq 1. \tag{1}$$

Two events are said to be mutually exclusive if they cannot both occur (under the complex of conditions S). For example, in throwing a die, the occurrence of an even number of spots and of a three are mutually exclusive. An event A is called the *union* of events A_1 and A_2 if it consists of the occurrence of at least one of the events A_1, A_2. For example, in throwing a die, the event A, consisting of rolling 1, 2, or 3, is the union of the events A_1 and A_2, where A_1 consists of rolling 1 or 2 and A_2 consists

[1] For brevity we now change $\mathbf{P}(A/S)$ to $\mathbf{P}(A)$.

of rolling 2 or 3. It is easy to see that for the number of occurrences m_1, m_2, and m of two mutually exclusive events A_1 and A_2 and their union $A = A_1 \cup A_2$, we have the equation $m = m_1 + m_2$, or for the corresponding frequencies $\nu = \nu_1 + \nu_2$.

This leads naturally to the following axiom for the addition of probabilities:

$$\mathbf{P}(A_1 \cup A_2) = \mathbf{P}(A_1) + \mathbf{P}(A_2), \tag{2}$$

if the events A_1 and A_2 are mutually exclusive and $A_1 \cup A_2$ denotes their union.

Further, for an event U which is certain, we naturally take

$$\mathbf{P}(U) = 1. \tag{3}$$

The whole mathematical theory of probability is constructed on the basis of simple axioms of the type (1), (2), and (3). From the point of view of pure mathematics, *probability* is a numerical function of "events," with a number of properties determined by axioms. The properties of probability, expressed by formulas (1), (2), and (3), serve as a sufficient basis for the construction of what is called the elementary theory of probability, if we do not insist on including in the axiomatization the concepts of an event itself, the union of events, and their intersection, as defined later. For the beginner it is more useful to confine himself to an intuitive understanding of the terms "event" and "probability," but to realize that although the meaning of these terms in practical life cannot be completely formalized, still this fact does not affect the complete formal precision of an axiomatized, purely mathematical presentation of the theory of probability.

The union of any given number of events $A_1, A_2, \dots, A_s$ is defined as the event A consisting of the occurrence of at least one of these events.

From the axiom of addition, we easily obtain for any number of pairwise mutually exclusive events $A_1, A_2, \dots, A_s$ and their union A,

$$\mathbf{P}(A) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_s)$$

(the so-called *theorem of the addition of probabilities*).

If the union of these events is an event that is certain (i.e., under the complex of conditions S one of the events A_k must occur), then

$$\mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_s) = 1.$$

In this case the system of events $A_1, \dots, A_s$ is called a *complete system* of events.

We now consider two events A and B, which, generally speaking, are not mutually exclusive. The event C is the intersection of the events A and B, written $C = AB$, if the event C consists of the occurrence of both A and B.[2]

For example, if the event A consists of obtaining an even number in the throw of a die and B consists of obtaining a multiple of three, then the event C consists of obtaining a six.

In a large number n of repeated trials, let the event A occur m times and the event B occur I times, in k of which B occurs together with the event A. The quotient k/m is called the conditional frequency of the event B under the condition A. The frequencies k/m, m/n, and k/n are connected by the formula

$$\frac{k}{m} = \frac{k}{n} : \frac{m}{n}$$

[2] Similarly, the intersection C of any number of events $A_1, A_2, \dots, A_s$ consists of the occurrence of all the given events.

which naturally gives rise to the following definition:

The conditional probability $\mathbf{P}(B \mid A)$ of the event B under the condition A is the quotient

$$\mathbf{P}(B \mid A) = \frac{\mathbf{P}(AB)}{\mathbf{P}(A)} .$$

Here it is assumed, of course, that $\mathbf{P}(A) \neq 0$.

If the events A and B are in no way essentially connected with each other, then it is natural to assume that event B will not appear more often, or less often, when A has occurred than when A has not occurred, i.e., that approximately $k/m \sim l/n$ or

$$\frac{k}{n} = \frac{k}{m}\,\frac{m}{n} \sim \frac{l}{n}\,\frac{m}{n} .$$

In this last approximate equation $m/n = \nu_A$ is the frequency of the event A, and $l/n = \nu_B$ is the frequency of the event B and finally $k/n = \nu_{AB}$ is the frequency of the intersection of the events A and B.

We see that these frequencies are connected by the relation

$$\nu_{AB} \sim \nu_A \, \nu_B.$$

For the probabilities of the events A, B, and AB, it is therefore natural to accept the corresponding exact equation

$$\mathbf{P}(AB) = \mathbf{P}(A) \cdot \mathbf{P}(B). \tag{4}$$

Equation (4) serves to define the *independence* of two events A and B.

Similarly, we may define the independence of any number of events. Also, we may give a definition of the independence of any number of experiments, which means, roughly speaking, that the outcome of any part of the experiments do not depend on the outcome of the rest.[3]

We now compute the probability P_k of precisely k occurrences of a certain event A in n independent tests, in each one of which the probability p of the occurrence of this event is the same. We denote by $\bar{A}$ the event that event A does not occur. It is obvious that

$$\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A) = 1 - p.$$

From the definition of the independence of experiments it is easy to see that the probability of any specific sequence consisting of k occurrences of A and $n-k$ nonoccurrences of A is equal to

$$p^k (1-p)^{n-k}. \tag{5}$$

[3] A more exact meaning of *independent experiments* is the following. We divide the n experiments in any way into two groups and let the event A consist of the result that all the experiments of the first group have certain preassigned outcomes, and the event B that the experiments of the second group have preassigned outcomes. The experiments are called independent (as a collection) if for arbitrary decomposition into two groups and arbitrarily preassigned outcomes the events A and B are independent in the sense of (4).

We will return in section 4 to a consideration of the objective meaning in the actual world of the independence of events.

Thus, for example, for $n = 5$ and $k = 2$ the probability of getting the sequence $A\bar{A}A\bar{A}\bar{A}$ will be $p(1-p)\,p(1-p)(1-p) = p^2(1-p)^3$.

By the theorem on the addition of probabilities, P_k will be equal to the sum of the probabilities of all sequences with k occurrences and $n-k$ nonoccurrences of the event A, i.e., P_k will be equal from (5) to the product of the number of such sequences by $p^k(1-p)^{n-k}$. The number of such sequences is obviously equal to the number of combinations of n things taken k at a time, since the k positive outcomes may occupy any k places in the sequence of n trials.

Finally we get

$$P_k = C_n^k p^k (1-p)^{n-k} \qquad (k = 0, 1, 2, \dots, n) \tag{6}$$

(which is called a binomial distribution).

In order to see how the definitions and formulas are applied, we consider an example that arises in the theory of gunfire.

Let five hits be sufficient for the destruction of the target. What interests us is the question whether we have the right to assume that 40 shots will insure the necessary five hits. A purely empirical solution of the problem would proceed as follows. For given dimensions of the target and for a given range, we carry out a large number (say 200) of firings, each consisting of 40 shots, and we determine how many of these firings produce at least five hits. If this result is achieved, for example, by 195 firings out of the 200, then the probability P is approximately equal to

$$P = \frac{195}{200} = 0.975.$$

If we proceed in this purely empirical way, we will use up 8,000 shells to solve a simple special problem. In practice, of course, no one proceeds in such a way. Instead, we begin the investigation by assuming that the scattering of the shells for a given range is independent of the size of the target. It turns out that the longitudinal and lateral deviations, from the mean point of landing of the shells, follow a law with respect to the frequency of deviations of various sizes that is illustrated in Figure 2.

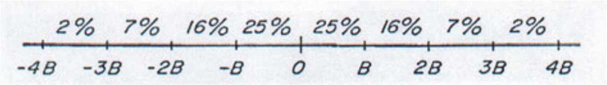

FIG. 2.

The letter B here denotes what is called the probable deviation. The probable deviation, generally speaking, is different for longitudinal and for lateral deviations and increases with increasing range. The probable deviations for different ranges for each type of gun and of shell are found empirically in firing practice on an artillery range. But the subsequent solution of all possible special problems of the kind described is carried out by calculations only.

For simplicity, we assume that the target has the form of a rectangle, one side of which is directed along the line of fire and has a length of two probable longitudinal deviations, while the other side is perpendicular to the line of fire and is equal in length to two probable lateral deviations. We assume further that the range has already been well established, so that the mean trajectory of the shells passes through its center (Figure 3).

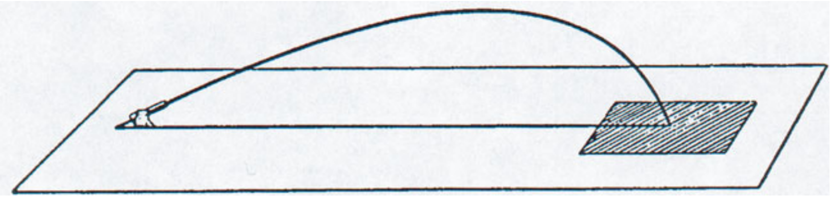

FIG. 3.

We also assume that the lateral and longitudinal deviations are independent.[4] Then for a given shell to fall on the target, it is necessary and sufficient that its longitudinal and lateral deviations do not exceed the corresponding probable deviations. From Figure 2 each of these events will be observed for about 50% of the shells fired, i.e., with probability $\frac{1}{2}$. The intersection of the two events will occur for about 25% of the shells fired; i.e., the probability that a specific shell will hit the target will be equal to

$$p = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

and the probability of a miss for a single shell will be

$$q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}.$$

Assuming that hits by the individual shells represent independent events, and applying the binomial formula (6), we find that the probability for getting exactly k hits in 40 shots will be

$$P_k = C_{40}^k p^k q^{40-k} = \frac{40 \cdot 39 \cdots (39-k)}{1 \cdot 2 \cdots k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{40-k}.$$

What concerns us is the probability of getting no less than five hits, and this is now expressed by the formula

$$P = \sum_{k=5}^{40} P_k.$$

But it is simpler to compute this probability from the formula $P = 1 - Q$, where

$$Q = \sum_{k=0}^{4} P_k$$

is the probability of getting less than five hits.

[4] This assumption of independence is borne out by experience.

We may calculate that

$$P_0 = \left(\frac{3}{4}\right)^{40} \sim 0.00001,$$

$$P_1 = 40\left(\frac{3}{4}\right)^{39}\frac{1}{4} \sim 0.00013,$$

$$P_2 = \frac{40\cdot 39}{2}\left(\frac{3}{4}\right)^{38}\left(\frac{1}{4}\right)^{2} \sim 0.00087,$$

$$P_3 = \frac{40\cdot 39\cdot 38}{2\cdot 3}\left(\frac{3}{4}\right)^{37}\left(\frac{1}{4}\right)^{3} \sim 0.0037,$$

$$P_4 = \frac{40\cdot 39\cdot 38\cdot 37}{2\cdot 3\cdot 4}\left(\frac{3}{4}\right)^{36}\left(\frac{1}{4}\right)^{4} \sim 0.0113,$$

so that

$$Q = 0.016, \qquad P = 0.984.$$

The probability P so obtained is somewhat closer to certainty than is usually taken to be sufficient in the theory of gunfire. Most often it is considered permissible to determine the number of shells needed to guarantee the result with probability 0.95.

The previous example is somewhat schematized, but it shows in sufficient detail the practical importance of probability calculations. After establishing by experiment the dependence of the probable deviations on the range (for which we did not need to fire a large number of shells), we were then able to obtain, by simple calculation, the answers to questions of the most diverse kind. The situation is the same in all other domains where the collective influence of a large number of random factors leads to a statistical law. Direct examination of the mass of observations makes clear only the very simplest statistical laws; it uncovers only a few of the basic probabilities involved. But then, by means of the laws of the theory of probability, we use these simplest probabilities to compute the probabilities of more complicated occurrences and deduce the statistical laws that govern them.

Sometimes we succeed in completely avoiding massive statistical material, since the probabilities may be defined by sufficiently convincing considerations of symmetry. For example, the traditional conclusion that a die, i.e., a cube made of a homogeneous material will fall, when thrown to a sufficient height, with equal probability on each of its faces was reached long before there was any systematic accumulation of data to verify it by observation. Systematic experiments of this kind have been carried out in the last three centuries, chiefly by authors of textbooks in the theory of probability, at a time when the theory of probability was already a well-developed science. The results of these experiments were satisfactory, but the question of extending them to analogous cases scarcely arouses interest. For example, as far as we know, no one has carried out sufficiently extensive experiments in tossing homogeneous dice with twelve sides. But there is no doubt that if we were to make 12,000 such tosses, the twelve-sided die would show each of its faces approximately a thousand times.

The basic probabilities derived from arguments of symmetry or homogeneity also play a large role in many serious scientific problems, for example in all problems of collision or near approach of molecules in random motion in a gas; another case where the successes have been equally great is the motion of stars in a galaxy. Of course, in these more delicate cases we prefer to check our theoretical assumptions by comparison with observation or experiment.

3. The law of large numbers and limit theorems. It is completely natural to wish for greater quantitative precision in the proposition that in a "long" series of tests the frequency of an occurrence comes "close" to its probability. But here we must form a clear notion of the delicate nature of the problem. In the most typical cases in the theory of probability, the situation is such that in an arbitrarily long series of tests it remains theoretically possible that we may obtain either of the two extremes for the value of the frequency

$$\frac{\mu}{n} = \frac{n}{n} = 1 \quad \text{and} \quad \frac{\mu}{n} = \frac{0}{n} = 0.$$

Thus, whatever may be the number of tests n, it is impossible to assert with complete certainty that we will have, say, the inequality

$$\left|\frac{\mu}{n} - p\right| < \frac{1}{10}\,.$$

For example, if the event A is the rolling of a six with a die, then in n trials, the probability that we will turn up a six on all n trials is $(\frac{1}{6})^n > 0$, in other words, with probability $(\frac{1}{6})^n$ we will obtain a frequency of rolling a six which is equal to *one*; and with probability $(1 - \frac{1}{6})^n > 0$ a six will not come up at all, i.e., the frequency of rolling a six will be equal to *zero*.

In all similar problems any nontrivial estimate of the closeness of the frequency to the probability cannot be made with complete certainty, but only with some probability less than one. For example, it may be shown that in independent tests,[5] with constant probability p of the occurrence of an event in each test the inequality

$$\left|\frac{\mu}{n} - p\right| < 0.02 \tag{7}$$

for the frequency μ/n will be satisfied, for $n = 10{,}000$ (and any p), with probability

$$P > 0.9999. \tag{8}$$

Here we wish first of all to emphasize that in this formulation the quantitative estimate of the closeness of the frequency μ/n to the probability p involves the introduction of a new probability P.

The practical meaning of the estimate (8) is this: If we carry out N sets of n tests each, and count the M sets in which inequality (7) is satisfied, then for sufficiently large N we will have approximately

$$\frac{M}{N} \approx P > 0.9999. \tag{9}$$

But if we wish to define the relation (9) more precisely, either with respect to the degree of closeness of M/N to P, or with respect to the confidence with which we may assert that (9) will be verified, then we must have recourse to general considerations of the kind introduced previously in discussing what is meant by the closeness of μ/n and p. Such considerations may be repeated as often as we like, but it is clear that this procedure will never allow us to be free of the necessity, at the last stage, of referring to probabilities in the primitive imprecise sense of this term.

It would be quite wrong to think that difficulties of this kind are peculiar in some way to the theory of probability. In a mathematical investigation of actual events, we always make a model of them. The discrepancies between the actual course of events and the theoretical model can, in its turn, be made the subject of mathematical

[5] The proof of the estimate (8) is discussed later in this section.

investigation. But for these discrepancies we must construct a model that we will use without formal mathematical analysis of the discrepancies which again would arise in it in actual experiment.

We note, moreover, that in an actual application of the estimate[6]

$$\mathbf{P}\left\{\left|\frac{\mu}{n}-p\right|<0.02\right\}>0.9999 \tag{10}$$

to one series of n tests we are already depending on certain considerations of symmetry: inequality (10) shows that for a very large number N of series of tests, relation (7) will be satisfied in no less than 99.99% of the cases; now it is natural to expect with great confidence that inequality (7) will apply in particular to that one of the sequence of n tests which is of interest to us, but we may expect this only if we have some reason for assuming that the position of this sequence among the others is a regular one, that is, that it has no special features.

The probabilities that we may decide to neglect are different in different practical situations. We noted earlier that our preliminary calculations for the expenditure of shells necessary to produce a given result meet the standard that the problem is to be solved with probability 0.95, i.e., that the neglected probabilities do not exceed 0.05. This standard is explained by the fact that if we were to make calculations neglecting a probability of only 0.01, let us say, we would necessarily require a much greater expenditure of shells, so that in practice we would conclude that the task could not be carried out in the time at our disposal, or with the given supply of shells.

In scientific investigations also, we are sometimes restricted to statistical methods calculated on the basis of neglecting probabilities of 0.05, although this practice should be adopted only in cases where the accumulation of more extensive data is very difficult. As an example of such a method let us consider the following problem. We assume that under specific conditions the customary medicine for treating a certain illness gives positive results 50% of the time, i.e., with probability 0.5. A new preparation is proposed, and to test its advantages we plan to use it in ten cases, chosen without bias from among the patients suffering from the illness. Here we agree that the advantage of the new preparation will be considered as proved if it gives a positive result in no less than eight cases out of the ten. It is easy to calculate that such a procedure involves the neglect of probabilities of the order of 0.05 of getting a wrong result, i.e., of indicating an advantage for the new preparation when in fact it is only equally effective or even worse than the old. For if in each of the ten experiments, the probability of a positive outcome is equal to p, then the probability of obtaining in ten experiments 10, 9, or 8 positive outcomes, is equal, respectively, to

$$P_{10}=p^{10},\quad P_9=10\,p^9(1-p),\quad P_8=45\,p^8(1-p)^2.$$

For the case $p=\frac{1}{2}$ the sum of these is

$$P=P_{10}+P_9+P_8=\frac{56}{1024}\sim 0.05.$$

In this way, under the assumption that in fact the new preparation is exactly as effective as the old, we risk with probability of order 0.05 the error of finding that the new preparation is better than the old. To reduce this probability to about 0.01, without increasing the number of experiments $n=10$, we will need to agree that the advantage of the new preparation is proved if it gives a positive result in no less than nine cases out of the ten. If this requirement seems too severe to the advocates of the

[6] This is the accepted notation for estimate (8) of the probability of inequality (7).

new preparation, it will be necessary to make the number of experiments considerably larger than 10. For example, for $n = 100$, if we agree that the advantage of the new preparation is proved for $\mu > 65$, then the probability of error will only be $P \approx 0.0015$.

For serious scientific investigations a standard of 0.05 is clearly insufficient; but even in such academic and circumstantial matters as the treatment of astronomical observations, it is customary to neglect probabilities of error of 0.001 or 0.003. On the other hand, some of the scientific results based on the laws of probability are considerably more reliable even than that; i.e., they involve the neglect of smaller probabilities. We will return to this question later.

In the previous examples, we have made use of particular cases of the binomial formula (6)

$$P_m = C_n^m p^m (1-p)^{n-m}$$

for the probability of getting exactly m positive results in n independent trials, in each one of which a positive outcome has probability p. Let us consider, by means of this formula, the question raised at the beginning of this section concerning the probability

$$P = \mathbf{P}\left\{ \left| \frac{\mu}{n} - p \right| < \varepsilon \right\}, \tag{11}$$

where μ is the actual number of positive results.[7] Obviously, this probability may be written as the sum of those P_m for which m satisfies the inequality

$$\left| \frac{m}{n} - p \right| < \varepsilon, \tag{12}$$

i.e., in the form

$$P = \sum_{m=m_1}^{m_2} P_m, \tag{13}$$

where m_1 is the smallest of the values of m satisfying inequality (12), and m_2 is the largest.

Formula (13) for fairly large n is hardly convenient for immediate calculation, a fact which explains the great importance of the asymptotic formula discovered by de Moivre for $p = \frac{1}{2}$ and by Laplace for general p. This formula allows us to find P_m very simply and to study its behavior for large n. The formula in question is

$$P_m \sim \frac{1}{\sqrt{2\pi np(1-p)}} \, e^{-(m-np)^2/[2np(1-p)]} \, . \tag{14}$$

If p is not too close to zero or one, it is sufficiently exact even for n to the order of 100. If we set

$$t = \frac{m - np}{\sqrt{np(1-p)}} \, , \tag{15}$$

then formula (14) becomes

$$P_m \sim \frac{1}{\sqrt{2\pi np(1-p)}} \, e^{-t^2/2} \, . \tag{16}$$

From (13) and (16) one may derive an approximate representation of the probability (11)

$$P \sim \frac{1}{\sqrt{2\pi}} \int_{-T}^{T} e^{-t^2/2} \, dt = F(T), \tag{17}$$

[7] Here, μ takes the values $m = 0, 1, \ldots, n$, with probability P_m; i.e., $\mathbf{P}\{\mu = m\} = P_m$.

where

$$T = \varepsilon\sqrt{\frac{n}{p(1-p)}}\,. \tag{18}$$

The difference between the left and right sides of (17) for fixed p, different from zero or one, approaches zero uniformly with respect to ε, as $n \to \infty$. For the function $F(T)$ detailed tables have been constructed. Here is a small excerpt from them:

T	1	2	3	4
F	0.68269	0.95450	0.99730	0.99993

.

For $T \to \infty$ the values of the function $F(T)$ converge to one.

From formula (17) we derive an estimate of the probability

$$\mathbf{P}\left\{\left|\frac{\mu}{n} - p\right| < 0.02\right\}$$

for $n = 10{,}000$. Since

$$T = \frac{2}{\sqrt{p(1-p)}},$$

we have

$$P \approx F\left(\frac{2}{\sqrt{p(1-p)}}\right).$$

Since the function $F(T)$ is monotonic increasing with increasing T, it follows for an estimate of P from the following which is independent of p, we must take the smallest possible (for the various p) value of T. Such a smallest value occurs for $p = \frac{1}{2}$ and is equal to 4. Thus, approximately

$$P \geqq F(4) = 0.99993. \tag{19}$$

In equality (19) no account is taken of the error arising from the approximate character of formula (17). By estimating the error involved here, we may show that in any case $P > 0.9999$.

In connection with this example of the application of formula (17), one should note that the estimates of the remainder term in formula (17) given in theoretical works on the theory of probability were for a long time unsatisfactory. Thus the applications of (17) and similar formulas to calculations based on small values of n, or with probabilities p very close to 0 or 1 (such probabilities are frequently of particular importance) were often based on experimental verification only of results of this kind for a restricted number of examples, and not on any valid estimates of the possible error. Also, it was shown by more detailed investigation that in many important practical cases the asymptotic formulas introduced previously require not only an estimate of the remainder term but also certain further refinements (without which the remainder term would be too large). In both directions the most complete results are due to S. N. Bernšteĭn.

Relations (11), (17), and (18) may be rewritten in the form

$$\mathbf{P}\left\{\left|\frac{\mu}{n} - p\right| < t\sqrt{\frac{p(1-p)}{n}}\right\} \sim F(t). \tag{20}$$

For sufficiently large t the right side of formula (20), which does not contain n, is arbitrarily close to one, i.e., to the value of the probability which gives complete certainty. We see, in this way, that, *as a rule, the deviation of the frequency* μ/n *from the probability* p *of order* $1/\sqrt{n}$. Such a proportionality between the exactness of a law of probability and the square root of the number of observations is typical for many other questions. Sometimes it is even said in popular simplifications that "the law of the square root of n" is the basic law of the theory of probability. Complete precision concerning this idea was attained through the introduction and systematic use by the great Russian mathematician P. L. Čebyšev of the concepts of "mathematical expectation" and "variance" for sums and arithmetic means of "random variables."

A *random variable* is the name given to a quantity which under given conditions S may take various values with specific probabilities. For us it is sufficient to consider random variables that may take on only a finite number of different values. To give the *probability distribution*, as it is called, of such a random variable ξ, it is sufficient to state its possible values $x_1, x_2, \dots, x_s$ and the probabilities

$$P_r = \mathbf{P}\{\xi = x_r\}.$$

The sum of these probabilities for all possible values of the variable ξ is always equal to one:

$$\sum_{r=1}^{s} P_r = 1.$$

The number investigated above of positive outcomes in n experiments may serve as an example of a random variable.

The *mathematical expectation* of the variable ξ is the expression

$$\mathbf{M}(\xi) = \sum_{r=1}^{s} P_r x_r,$$

and the *variance* of ξ is the mathematical expectation of the square of the deviation $\xi - \mathbf{M}(\xi)$, i.e., the expression

$$\mathbf{D}(\xi) = \sum_{r=1}^{s} P_r \big[x_r - \mathbf{M}(\xi)\big]^2.$$

The square root of the variance

$$\sigma_\xi = \sqrt{\mathbf{D}(\xi)}$$

is called the *standard deviation* (of the variable from its mathematical expectation $\mathbf{M}(\xi)$).

At the basis of the simplest applications of variance and standard deviation lies the famous inequality of Čebyšev

$$\mathbf{P}\big\{\big|\xi - \mathbf{M}(\xi)\big| \leqq t\sigma_\xi\big\} \geqq 1 - \frac{1}{t^2}. \tag{21}$$

It shows that deviations of ξ from $\mathbf{M}(\xi)$ significantly greater than σ_ξ are rare.

As for the sum of random variables

$$\xi = \xi^{(1)} + \xi^{(2)} + \cdots + \xi^{(n)},$$

their mathematical expectations always satisfy the equation

$$\mathbf{M}(\xi) = \mathbf{M}(\xi^{(1)}) + \mathbf{M}(\xi^{(2)}) + \cdots + \mathbf{M}(\xi^{(n)}). \tag{22}$$

But the analogous equation for the variance

$$\mathbf{D}(\xi) = \mathbf{D}(\xi^{(1)}) + \mathbf{D}(\xi^{(2)}) + \cdots + \mathbf{D}(\xi^{(n)}) \tag{23}$$

is true only under certain restrictions. For the validity of equation (23) it is sufficient, for example, that the variables $\xi^{(i)}$ and $\xi^{(j)}$ with different indices not be "correlated" with one another, i.e., that for $i \neq j$ the equation[8]

$$\mathbf{M}\Big\{\big[\xi^{(i)} - \mathbf{M}(\xi^{(i)})\big]\,\big[\xi^{(j)} - \mathbf{M}(\xi^{(j)})\big]\Big\} = 0 \tag{24}$$

be satisfied.

In particular, equation (24) holds if the variables $\xi^{(i)}$ and $\xi^{(j)}$ are independent of each other.[9] Consequently, for mutually independent terms equation (23) always holds. For the arithmetic mean

$$\zeta = \frac{1}{n}\left(\xi^{(1)} + \xi^{(2)} + \cdots + \xi^{(n)}\right)$$

it follows from (23) that

$$\mathbf{D}(\zeta) = \frac{1}{n^2}\left[\mathbf{D}(\xi^{(1)}) + \mathbf{D}(\xi^{(2)}) + \cdots + \mathbf{D}(\xi^{(n)})\right]. \tag{25}$$

We now assume that for each of these terms the variance does not exceed a certain constant

$$\mathbf{D}(\xi^{(i)}) \leqq C^2.$$

Then from (25)

$$\mathbf{D}(\zeta) \leqq \frac{C^2}{n},$$

and from Čebyšev's inequality for any t

$$\mathbf{P}\left\{\left|\zeta - \mathbf{M}(\zeta)\right| \leqq \frac{tC}{\sqrt{n}}\right\} \geqq 1 - \frac{1}{t^2}. \tag{26}$$

Inequality (26) expresses what is called the law of large numbers, in the form established by Čebyšev: If the variables $\xi^{(i)}$ are mutually independent and have bounded variance, then for increasing n the arithmetic mean ζ will deviate more and more rarely from the mathematical expectation $\mathbf{M}(\zeta)$.

More precisely, the *sequence of variables*

$$\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)}, \ldots$$

[8] The *correlation coefficient* between the variables $\xi^{(i)}$ and $\xi^{(j)}$ is the expression

$$R = \frac{\mathbf{M}\{[\xi^{(i)} - \mathbf{M}(\xi^{(i)})][\xi^{(j)} - \mathbf{M}(\xi^{(j)})]\}}{\sigma_{\xi^{(i)}}\sigma_{\xi^{(j)}}}.$$

If $\sigma_{\xi^{(i)}} > 0$ and $\sigma_{\xi^{(j)}} > 0$, then condition (24) is equivalent to saying that $R = 0$.

The correlation coefficient R characterizes the degree of dependence between random variables. $|R| \leqq 1$ always, and $R = \pm 1$ only for a linear relationship

$$\eta = a\xi + b \qquad (a \neq 0).$$

For independent variables $R = 0$.

[9] The independence of two random variables ξ and η, which may assume, respectively, the values $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$, is defined to mean that for any i and j the events $A_i = \{\xi = x_i\}$ and $B_j = \{\eta = y_j\}$ are independent in the sense of the definition given in section 2.

is said to obey the law of large numbers if for the corresponding arithmetic means ζ *and for any constant* $\varepsilon > 0$

$$\mathbf{P}\Big\{\big|\zeta - \mathbf{M}(\zeta)\big| \leqq \varepsilon\Big\} \to 1 \tag{27}$$

for $n \to \infty$.

In order to pass from inequality (26) to the limiting relation (27) it is sufficient to put

$$t = \varepsilon\,\frac{\sqrt{n}}{C}\,.$$

A large number of investigations of A. A. Markov, S. N. Bernšteĭn, A. Ja. Hinčin, and others were devoted to the question of widening as far as possible the conditions under which the limit relation (27) is valid, i.e., the conditions for the validity of the law of large numbers. These investigations are of basic theoretical significance, but still more important is an exact study of the probability distribution for the variable $\zeta - \mathbf{M}(\zeta)$.

One of the greatest services rendered by the classical Russian school of mathematicians to the theory of probability is the establishment of the fact that under very wide conditions the equation

$$\mathbf{P}\big\{t_1\sigma_\zeta < \zeta - \mathbf{M}(\zeta) < t_2\sigma_\zeta\big\} \sim \frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-t^2/2}\,dt \tag{28}$$

is asymptotically valid (i.e., with greater and greater exactness as n increases beyond all bounds).

Čebyšev gave an almost complete proof of this formula for the case of independent and bounded terms. Markov closed a gap in Čebyšev's argument and widened the conditions of applicability of formula (28). Still more general conditions were given by Ljapunov. The applicability of formula (28) to the sum of mutually dependent terms was studied with particular completeness by S. N. Bernšteĭn.

Formula (28) embraces such a large number of particular cases that it has long been called the central limit theorem in the theory of probability. Even though it has been shown lately to be included in a series of more general laws its value can scarcely be overrated even at the present time.

If the terms are independent and their variances are all the same, and are equal to

$$\mathbf{D}(\xi^{(i)}) = \sigma^2,$$

then it is convenient, using relation (25), to put formula (28) into the form

$$\mathbf{P}\left\{\frac{t_1\sigma}{\sqrt{n}} < \zeta - \mathbf{M}(\zeta) < \frac{t_2\sigma}{\sqrt{n}}\right\} \sim \frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-t^2/2}\,dt. \tag{29}$$

Let us show that relation (29) contains the solution of the problem, considered earlier, of evaluating the deviation of the frequency μ/n from the probability p. For this we introduce the random variables $\xi^{(i)}$, defined as follows:

$$\xi^{(i)} = \begin{cases} 0, & \text{if the } i\text{th test has a negative outcome,}\\ 1, & \text{if the } i\text{th test has a positive outcome.}\end{cases}$$

It is easy to verify that then

$$\mu = \xi^{(1)} + \xi^{(2)} + \cdots + \xi^{(n)}, \qquad \frac{\mu}{n} = \zeta,$$

$$\mathbf{M}(\xi^{(i)}) = p, \quad \mathbf{D}(\xi^{(i)}) = p(1-p), \quad \mathbf{M}(\zeta) = p,$$

and formula (29) gives

$$\mathbf{P}\left\{t_1\sqrt{\frac{p(1-p)}{n}} < \frac{\mu}{n} - p < t_2\sqrt{\frac{p(1-p)}{n}}\right\} \sim \frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-t^2/2}\,dt,$$

which for $t_1 = -t$, $t_2 = t$ leads again to formula (20).

4. Further remarks on the basic concepts of the theory of probability. In speaking of random events, which have the property that their frequencies tend to become stable, i.e., in a long sequence of experiments repeated under fixed conditions, their frequencies are grouped around some *standard level*, called their probability $\mathbf{P}(A/S)$, we were guilty, in section 1, of a certain vagueness in our formulations, in two respects. In the first place, we did not indicate how long the sequence of experiments n_r must be in order to exhibit beyond all doubt the existence of the supposed stability; in other words, we did not say what deviations of the frequencies μ_r/n_r from one another or from their standard level p were allowable for sequences of trials $n_1, n_2, \dots, n_s$ of given length. This inexactness in the first stage of formulating the concepts of a new science is unavoidable. It is no greater than the well-known vagueness surrounding the simplest geometric concepts of point and straight line and their *physical* meaning. This aspect of the matter was made clear in section 3.

More fundamental, however, is the second lack of clearness concealed in our formulations; it concerns the manner of forming the sequences of trials in which we are to examine the stability of the frequency of occurrence of the event A.

As stated earlier, we are led to statistical and probabilistic methods of investigation in those cases in which an exact specific prediction of the course of events is impossible. But if we wish to create in some artificial way a sequence of events that will be, as far as possible, purely random, then we must take special care that there shall be no methods available for determining in advance those cases in which A is likely to occur with more than normal frequency.

Such precautions are taken, for example, in the organization of government lotteries. If in a given lottery there are to be M winning tickets in a drawing of N tickets, then the probability of winning for an individual ticket is equal to $p = M/N$. This means that in whatever manner we select, in advance of the drawing, a sufficiently large set of n tickets, we can be practically certain that the ratio μ/n of the number μ of winning tickets in the chosen set to the whole number n of tickets in this set will be close to p. For example, people who prefer tickets labelled with an even number will not have any systematic advantage over those who prefer tickets labelled with odd numbers, and in exactly the same way there will be no advantage in proceeding on the principle, say, that it is always better to buy tickets with numbers having exactly three prime factors, or tickets whose numbers are close to those that were winners in the preceding lottery, etc.

Similarly, when we are firing a well-constructed gun of a given type, with a well-trained crew and with shells that have been subjected to a standard quality control, the deviation from the mean position of the points of impact of the shells will be less than the previously determined probable deviation B in approximately *half* the cases. This fraction remains the same in a series of successive trials, and also in case we count separately the number of deviations that are less than B for even-numbered shots (in the order of firing) or for odd-numbered. But it is completely possible that if we were to make a selection of particularly homogeneous shells (with respect to weight, etc.), the scattering would be considerably decreased, i.e., we would have a sequence of firings for which the fraction of the deviations which are greater than the standard B would be considerably less than a half.

Thus, to say that an event A is "random" or "stochastic" and to assign it a definite probability

$$p = \mathbf{P}(A/S)$$

is possible only when we have already determined the class of allowable ways of setting up the series of experiments. The nature of this class will be assumed to be included in the conditions S.

For *given* conditions S the properties of the event A of being random and of having the probability $p = \mathbf{P}(A/S)$ express the objective character of the connection between the condition S and the event A. In other words, there exists no event which is absolutely random; an event is random or is predetermined depending on the connection in which it is considered, but under specific conditions an event may be random in a completely nonsubjective sense, i.e., independently of the state of knowledge of any observer. If we imagine an observer who can master all the detailed distinctive properties and particular circumstances of the flight of shells, and can thus predict for each one of them the deviation from the mean trajectory, his presence would still not prevent the shells from scattering in accordance with the laws of the theory of probability, provided, of course, that the shooting was done in the usual manner, and not according to instructions from our imaginary observer.

In this connection we note that the formation of a series of the kind discussed earlier, in which there is a tendency for the frequencies to become constant in the sense of being grouped around a normal value, namely the probability, proceeds in the actual world in a manner completely independent of our intervention. For example, it is precisely by virtue of the random character of the motion of the molecules in a gas that the number of molecules which, even in a very small interval of time, strike an arbitrarily preassigned small section of the wall of the container (or of the surface of bodies situated in the gas) proves to be proportional with very great exactness to the area of this small piece of the wall and to the length of the interval of time. Deviations from this proportionality in cases where the number of hits is not large also follow the laws of the theory of probability and produce phenomena of the type of Brownian motion, of which more will be said later.

We turn now to the objective meaning of the concept of independence. We recall that the conditional probability of an event A under the condition B is defined by the formula

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)}. \tag{30}$$

We also recall that events A and B are called independent if, as in (4),

$$\mathbf{P}(AB) = \mathbf{P}(A)\,\mathbf{P}(B).$$

From the independence of the events A and B and the fact that $P(B) > 0$ it follows that

$$\mathbf{P}(A \mid B) = \mathbf{P}(A).$$

All the theorems of the mathematical theory of probability that deal with independent events apply to any events satisfying the condition (4), or to its generalization to the case of the mutual independence of several events. These theorems will be of little interest, however, if this definition bears no relation to the properties of objective events which are independent in the causal sense.

It is known, for example, that the probability of giving birth to a boy is, with sufficient stability, $\mathbf{P}(A) = 22/43$. If B denotes the condition that the birth occur on a day of the conjunction of Jupiter with Mars, then under the assumption that the position of the planets does not influence the fate of individuals, the conditional probability $\mathbf{P}(A \mid B)$ has the same value: $\mathbf{P}(A \mid B) = 22/43$; i.e., the actual calculation of the frequency of births of boys under such special astrological conditions would give just the same frequency 22/43. Although such a calculation has probably never been carried out on a sufficiently large scale, still there is no reason to doubt what the result would be.

We give this example, from a somewhat outmoded subject, in order to show that the development of human knowledge consists not only in establishing valid relations among phenomena, but also in refuting imagined relations, i.e., in establishing in relevant cases the thesis of the independence of any two sets of events. This unmasking of the meaningless attempts of the astrologers to connect two sets of events that are not in fact connected is one of the classic examples.

Naturally, in dealing with the concept of independence, we must not proceed in too absolute a fashion. For example, from the law of universal gravitation, it is an undoubted fact that the motions of the moons of Jupiter have a certain effect, say, on the flight of an artillery shell. But it is also obvious that in practice this influence may be ignored. From the philosophical point of view, we may perhaps, in a given concrete situation, speak more properly not of the independence but of the insignificance of the dependence of certain events. However that may be, the independence of events in the cited concrete and relative sense of this term in no way contradicts the principle of the universal interconnection of all phenomena; it serves only as a necessary supplement to this principle.

The computation of probabilities from formulas derived by assuming the independence of certain events is still of practical interest in cases where the events were originally independent but became interdependent as a result of the events themselves. For example, one may compute probabilities for the collision of particles of cosmic radiation with particles of the medium penetrated by the radiation, on the assumption that the motion of the particles of the medium, up to the time of the appearance near them of a rapidly moving particle of cosmic radiation, proceeds independently of the motion of the cosmic particle. One may compute the probability that a hostile bullet will strike the blade of a rotating propeller, on the assumption that the position of the blade with respect to the axis of rotation does not depend on the trajectory of the bullet, a supposition that will of course be wrong with respect to the bullets of the aviator himself, since they are fired between the blades of the rotating propeller. The number of such examples may be extended without limit.

It may even be said that wherever probabilistic laws turn up in any clear-cut way we are dealing with the influence of a large number of factors that, if not entirely independent of one another, are interconnected only in some weak sense.

This does not at all mean that we should uncritically introduce assumptions of independence. On the contrary, it leads us, in the first place, to be particularly careful in the choice of criteria for testing hypotheses of independence, and second, to be very careful in investigating the borderline cases where dependence between the facts must be assumed but is of such a kind as to introduce complications into the relevant laws of probability. We noted earlier that the classical Russian school of the theory of probability has carried out far-reaching investigations in this direction.

To bring to an end our discussion of the concept of independence, we note that, just as with the definition of independence of two events given in formula (4), the formal definition of the independence of several random variables is considerably broader than the concept of independence in the practical world, i.e., the absence of causal connection.

Let us assume, for example, that the point ξ falls in the interval $[0, 1]$ in such a manner for

$$0 \leqq a \leqq b \leqq 1$$

the probability that it belongs to the segment $[a, b]$ is equal to the length of this segment $b - a$. It is easy to prove that in the expansion

$$\xi = \frac{\alpha_1}{10} + \frac{\alpha_2}{100} + \frac{\alpha_3}{1000} + \cdots$$

of the abscissa of the point ξ in a decimal fraction, the digits α_k will be mutually independent, although they are interconnected by the way they are produced.[10] (From this fact follow many theoretical results, some of which are of practical interest.)

Such flexibility in the formal definition of independence should not be considered as a blemish. On the contrary it merely extends the domain of applicability of theorems established for one or another assumption of independence. These theorems are equally applicable in cases where the independence is postulated on the basis of practical considerations and in cases where the independence is proved by computation proceeding from previous assumptions concerning the probability distributions of the events and the random variables under study.

In general, investigation of the formal structure of the mathematical apparatus of the theory of probability has led to interesting results. It turns out that this apparatus occupies a very definite and clear-cut place in the classification, which nowadays is gradually becoming clear in outline, of the basic objects of study in contemporary mathematics.

We have already spoken of the concepts of intersection AB and union $A \cup B$ of the events A and B. We recall that events are called mutually exclusive if their intersection is empty, i.e., if $AB = N$, where N is the symbol for an impossible event.

The basic axiom of the elementary theory of probability consists of the requirement (cf. section 2) that under the condition $AB = N$ we have the equation

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

The basic concepts of the theory of probability, namely random events and their probabilities, are completely analogous in their properties to plane figures and their areas. It is sufficient to understand by AB the intersection (common part) of two figures, by $A \cup B$ their union, by N the conventional "empty" figure, and by $\mathbf{P}(A)$ the area of the figure A, whereupon the analogy is complete.

The same remarks apply to the volumes of three-dimensional figures.

The most general theory of entities of such a type, which contains as special cases the theory of volume and area, is now usually called *measure theory*, discussed in Chapter XV (vol. 3)[11] in connection with the theory of functions of a real variable.

[10] This is also valid, for any n, for the digits α_k in the expansion of the number ξ in the fraction

$$\xi = \frac{\alpha_1}{n} + \frac{\alpha_2}{n^2} + \frac{\alpha_3}{n^3} + \cdots .$$

[11] Editor's note: See A. D. Aleksandrov, A. N. Kolmogorov, and M. A. Lavrentiev, eds., *Mathematics: Its Content, Methods, and Meaning*, 3 vols., MIT Press, Cambridge, MA, 1969.

It remains only to note that in the theory of probability, in comparison with the general theory of measure or in particular with the theory of area and volume, there is a certain special feature: A probability is never greater than one. This maximal probability holds for a necessary event U.

$$\mathbf{P}(U) = 1.$$

The analogy is by no means superficial. It turns out that the whole mathematical theory of probability from the formal point of view may be constructed as a theory of measure, making the special assumption that the measure of "the entire space" U is equal to one.[12]

Such an approach to the matter has produced complete clarity in the formal construction of the mathematical theory of probability and has also led to concrete progress not only in this theory itself but in other theories closely related to it in their formal structure. In the theory of probability success has been achieved by refined methods developed in the metric theory of functions of a real variable and at the same time probabilistic methods have proved to be applicable to questions in neighboring domains of mathematics not "by analogy," but by a formal and strict transfer of them to the new domain. Wherever we can show that the axioms of the theory of probability are satisfied, the results of these axioms are applicable, even though the given domain has nothing to do with randomness in the actual world.

The existence of an axiomatized theory of probability preserves us from the temptation "to define" probability by methods that claim to construct a strict, purely formal mathematical theory on the basis of features of probability that are immediately suggested by the natural sciences. Such definitions roughly correspond to the "definition" in geometry of a point as the result of trimming down a physical body an infinite number of times, each time decreasing its diameter by a factor of 2.

With definitions of this sort, probability is taken to be the limit of the frequency as the number of experiments increases beyond all bounds. The very assumption that the experiments are probabilistic, i.e., that the frequencies tend to cluster around a constant value, will remain valid (and the same is true for the "randomness" of any particular event) only if certain conditions are kept fixed for an unlimited time and with absolute exactness. Thus the exact passage to the limit

$$\frac{\mu}{n} \to p$$

cannot have any objective meaning. Formulation of the principle of stability of the frequencies in such a limit process demands that we define the allowable methods of setting up an infinite sequence of experiments, and this can only be done by a mathematical fiction. This whole conglomeration of concepts might deserve serious consideration if the final result were a theory of such distinctive nature that no other means existed of putting it on a rigorous basis. But, as was stated earlier, the mathematical theory of probability may be based on the theory of measure, in its present-day form, by simply adding the condition

$$\mathbf{P}(U) = 1.$$

In general, for any practical analysis of the concept of probability, there is no need to refer to its formal definition. It is obvious that concerning the purely formal side

[12] Nevertheless, because of the nature of its problems, the theory of probability remains an independent mathematical discipline; its basic results (presented in detail in section 3) appear artificial and unnecessary from the point of view of pure measure theory.

of probability, we can only say the following: The probability $\mathbf{P}(A/S)$ is a number around which, under conditions S determining the allowable manner of setting up the experiments, the frequencies have a tendency to be grouped, and that this tendency will occur with greater and greater exactness as the experiments, always conducted in such a way as to preserve the original conditions, become more numerous, and finally that the tendency will reach a satisfactory degree of reliability and exactness during the course of a practicable number of experiments.

In fact, the problem of importance, in practice, is not to give a formally precise definition of randomness but to clarify as widely as possible the conditions under which randomness of the cited type will occur. One must clearly understand that, in reality, hypotheses concerning the probabilistic character of any phenomenon are very rarely based on immediate statistical verification. Only in the first stage of the penetration of probabilistic methods into a new domain of science has the work consisted of purely empirical observation of the constancy of frequencies. From section 3, we see that statistical verification of the constancy of frequencies with an exactness of ε requires a series of experiments, each consisting of $n = 1/\varepsilon^2$ trials. For example, in order to establish that in a given concrete problem the probability is defined with an exactness of 0.0001, it is necessary to carry out a series of experiments containing approximately 100,000,000 trials in each.

The hypothesis of probabilistic randomness is much more often introduced from considerations of symmetry or of successive series of events, with subsequent verification of the hypothesis in some indirect way. For example, since the number of molecules in a finite volume of gas is of the order of 10^{20} or more, the number $\sqrt{n}$, corresponding to the probabilistic deductions made in the kinetic theory of gases, is very large, so that many of these deductions are verified with great exactness. Thus, the pressures on the opposite sides of a plate suspended in still air, even if the plate is of microscopic dimensions, turn out exactly the same, although an excess of pressure on one side of the order of a thousandth of one percent can be detected in a properly arranged experiment.

5. Deterministic and random processes. The principle of causal relation among phenomena finds its simplest mathematical expression in the study of physical processes by means of differential equations as demonstrated in a series of examples in section 1 of Chapter V.

Let the state of the system under study be defined at the instant of time t by n parameters

$$x_1, x_2, \dots, x_n.$$

The rates of change of these parameters are expressed by their derivatives with respect to time

$$\dot{x}_k = \frac{dx_k}{dt}\,.$$

If it is assumed that these rates are functions of the values of the parameters, then we get a system of differential equations

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, \dots, x_n), \\ \dot{x}_2 &= f_2(x_1, x_2, \dots, x_n), \\ &\dots\dots\dots\dots\dots\dots \\ \dot{x}_n &= f_n(x_1, x_2, \dots, x_n). \end{aligned}$$

The greater part of the laws of nature discovered at the time of the birth of mathematical physics, beginning with Galileo's law for falling bodies, are expressed in just such a manner. Galileo could not express his discovery in this standard form, since in his time the corresponding mathematical concepts had not yet been developed, and this was first done by Newton.

In mechanics and in any other fields of physics, it is customary to express these laws by differential equations of the second order. But no new principles are involved here; for if we denote the rates $\dot{x}_k$ by the new symbols

$$v_k = \dot{x}_k,$$

we get for the second derivative of the quantities x_k the expressions

$$\frac{d^2x_k}{dt^2} = \dot{v}_k,$$

and the equations of the second order for the n quantities $x_1, x_2, \dots, x_n$ become equations of the first order for the $2n$ quantities $x_1, \dots, x_n$, $v_1, v_2, \dots, v_n$.

As an example, let us consider the fall of a heavy body in the atmosphere of the earth. If we consider only short distances above the surface, we may assume that the resistance of the medium depends only on the velocity and not on the height. The state of the system under study is characterized by two parameters: the distance z of the body from the surface of the earth, and its velocity v. The change of these two quantities with time is defined by the two differential equations

$$\begin{aligned} \dot{z} &= -v, \\ \dot{v} &= g - f(v), \end{aligned} \tag{31}$$

where g is the acceleration of gravity and $f(v)$ is some "law of resistance" for the given body.

If the velocity is not great and the body is sufficiently massive, say a stone of moderate size falling from a height of several meters, the resistance of the air may be neglected and equations (31) are transformed into the equations

$$\begin{aligned} \dot{z} &= -v, \\ \dot{v} &= g. \end{aligned} \tag{32}$$

If it is assumed that at the initial instant of time t_0 the quantities z and v have values z_0 and v_0, then it is easy to solve equations (32) to obtain the formula

$$z = z_0 - v(t - t_0) - g\left(\frac{t - t_0}{2}\right)^2,$$

which describes the whole process of falling. For example, if $t_0 = 0$, $v_0 = 0$ we get

$$z = z_0 - \frac{gt^2}{2},$$

found by Galileo.

In the general case, the integration of equations (31) is more difficult, although the basic result, with very general restrictions on the function $f(v)$, remains the same: Given the values z_0 and v_0 at the initial instant t_0, the values of z and v for all further instants t are computed uniquely, up to the time that the falling body hits the surface of the earth. Theoretically, this last restriction may also be removed, if we assume that the fall is extended to negative values of z. For problems set up in this manner, the following may be established: If the function $f(v)$ is monotone for increasing v and tends to infinity for $v \to \infty$, then if the fall continues unchecked, i.e., for unbounded growth of the variable t, the velocity v tends to a constant limiting value c, which is the solution of the equation

$$g = f(c).$$

From the intuitive point of view, this result of the mathematical analysis of the problem is quite understandable: The velocity of fall increases up to the time that the accelerative force of gravity is balanced by the resistance of the air. For a jump with an open parachute, the stationary velocity v of about five meters per second is attained rather quickly.[13] For a long jump with unopened parachute the resistance of the air is less, so that the stationary velocity is greater and is attained only after the parachutist has fallen a very long way.

For the falling of light bodies like a feather tossed into the air or a bit of fluff, the initial period of acceleration is very short, often quite unobservable. The stationary rate of falling is established very quickly, and to a standard approximation we may consider that throughout the fall $v = c$. In this case we have only one differential equation

$$\dot{z} = -c,$$

which is integrated very simply:

$$z = z_0 - c(t - t_0).$$

This is how a bit of fluff will fall in perfectly still air.

This deterministic conception is treated in a completely general way in the contemporary theory of dynamical systems, to which is dedicated a series of important works by Soviet mathematicians, N. N. Bogoljubov, V. V. Stepanov, and many others. This general theory also includes as special cases the mathematical formulation of physical phenomena in which the state of a system is not defined by a finite number of parameters as in the earlier case, but by one or more functions, for example, in the mechanics of continuous media. In such cases the elementary laws for change of state in "infinitely small" intervals of time are given not by ordinary but by partial differential equations or by some other means. But the features common to all deterministic mathematical formulations of actual processes are: first, that the state of the system under study is considered to be completely defined by some mathematical entity ω (a set of n real numbers, one or more functions, and so forth); and second, that the later values for instants of time $t > t_0$ are uniquely determined by the value ω_0 at the initial instant t_0

$$\omega = F(t_0, \omega_0, t).$$

For phenomena described by differential equations the process of finding the function ϕ consists, as we have seen, in integrating these differential equations with the initial conditions $\omega = \omega_0$ for $t = t_0$.

The proponents of mechanistic materialism assumed that such a formulation is an exact and direct expression of the deterministic character of the actual phenomena, of the physical principle of causation. According to Laplace, the state of the world at a given instant is defined by an infinite number of parameters, subject to an infinite

[13] This statement is to be taken in the sense that in practice v soon gets quite close to c.

number of differential equations. If some "universal mind" could write down all these equations and integrate them, it could then predict with complete exactness, according to Laplace, the entire evolution of the world in the infinite future.

But in fact this quantitative mathematical infinity is extremely coarse in comparison with the qualitatively inexhaustible character of the real world. Neither the introduction of an infinite number of parameters nor the description of the state of continuous media by functions of a point in space is adequate to represent the infinite complexity of actual events.

As was emphasized in section 3 of Chapter V, the study of actual events does not always proceed in the direction of increasing the number of parameters introduced into the problem; in general, it is far from expedient to complicate the ω which describes the separate "states of the system" in our mathematical scheme. The art of the investigation consists rather in finding a very simple space Ω (i.e., a set of values of ω or in other words, of different possible states of the system),[14] such that if we replace the actual process by varying the point ω in a determinate way over this space, we can include all the *essential* aspects of the actual process.

But if from an actual process we abstract its essential aspects, we are left with a certain residue which we must consider to be random. The neglected random factors always exercise a certain influence on the course of the process. Very few of the phenomena that admit mathematical investigation fail, when theory is compared with observation, to show the influence of ignored random factors. This is more or less the state of affairs in the theory of planetary motion under the force of gravity: The distance between planets is so large in comparison with their size that the idealized representation of them as material points is almost perfectly satisfactory; the space in which they are moving is filled with such dispersed material that its resistance to their motion is vanishingly small; the masses of the planets are so large that the pressure of light plays almost no role in their motions. These exceptional circumstances explain the fact that the mathematical solution for the motion of a system of n material points, whose "states" are described by $6n$ parameters[15] which take into account only the force of gravity, agrees so astonishingly well with observation of the motion of the planets.

Somewhat similar to the case of planetary motion is the flight of an artillery shell under gravity and resistance of the air. This is also one of the classical regions in which mathematical methods of investigation were comparatively easy and quickly produced great success. But here the role of the perturbing random factors is significantly larger and the scattering of the shells, i.e., their deviation from the theoretical trajectory reaches tens of meters, or for long ranges even hundreds of meters. These deviations are caused partly by random deviations in the initial direction and velocity, partly by random deviations in the mass and the coefficient of resistance of the shell, and partly by gusts and other irregularities in the wind and the other random factors governing the extraordinarily complicated and changing conditions in the actual atmosphere of the earth.

The scattering of shells is studied in detail by the methods of the theory of probability, and the results of this study are essential for the practice of gunnery.

But what does it mean, properly speaking, to study random events? It would seem that, when the random "residue" for a given formulation of a phenomenon proves to

[14] In the example given earlier of a falling body, the phase space is the system of pairs of numbers (z, v), i.e., a plane. For phase spaces in general, see Chapters XVII and XVIII.

[15] The three coordinates and the three components of the velocity of each point.

be so large that it can not be neglected, then the only possible way to proceed is to describe the phenomenon more accurately by introducing new parameters and to make a more detailed study by the same method as before.

But in many cases such a procedure is not realizable in practice. For example, in studying the fall of a material body in the atmosphere, with account taken of an irregular and gusty (or, as one usually says, turbulent) wind flow, we would be required to introduce, in place of the two parameters z and v, an altogether unwieldy mathematical apparatus to describe this structure completely.

But in fact this complicated procedure is necessary only in those cases where for some reason we must determine the influence of these residual "random" factors in all detail and separately for each individual factor. Fortunately, our practical requirements are usually quite different; we need only estimate the total effect exerted by the random factors for a long interval of time or for a large number of repetitions of the process under study.

As an example, let us consider the shifting of sand in the bed of a river, or in a hydroelectric construction. Usually this shifting occurs in such a way that the greater part of the sand remains undisturbed, while only now and then a particularly strong turbulence near the bottom picks up individual grains and carries them to a considerable distance, where they are suddenly deposited in a new position. The purely theoretical motion of each grain may be computed individually by the laws of hydrodynamics, but for this it would be necessary to determine the initial state of the bottom and of the flow in every detail and to compute the flow step by step, noting those instants when the pressure on any particular grain of sand becomes sufficient to set it in motion, and tracing this motion until it suddenly comes to an end. The absurdity of setting up such a problem for actual scientific study is obvious. Nevertheless the average laws or, as they are usually called, the statistical laws of shifting of sand over river bottoms are completely amenable to investigation.

Examples of this sort, where the effect of a large number of random factors leads to a completely clear-cut statistical law, could easily be multiplied. One of the best known and at the same time most fascinating of these, in view of the breadth of its applications, is the kinetic theory of gases, which shows how the joint influence of random collisions of molecules gives rise to exact laws governing the pressure of a gas on the wall, the diffusion of one gas through another, and so forth.

6. Random processes of Markov type. To A. A. Markov is due the construction of a probabilistic scheme which is an immediate generalization of the deterministic scheme of section 5 described by the equation

$$\omega = F(t_0, \omega_0, t).$$

It is true that Markov considered only the case where the phase space of the system consists of a finite number of states $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ and studied the change of state of the system only for changes of time t in discrete steps. But in this extremely schematic model he succeeded in establishing a series of fundamental laws.

Instead of a function F, uniquely defining the state ω at time $t > t_0$ corresponding to the state ω_0 at time t_0, Markov introduced the probabilities

$$\mathbf{P}(t_0, \omega_i;\, t, \omega_j)$$

of obtaining the state ω_j at time t under the condition that at time t_0 we had the state ω_i. These probabilities are connected for any three instants of time

$$t_0 < t_1 < t_2$$

by a relation, introduced by Markov, which may be called the basic equation for a Markov process

$$\mathbf{P}(t_0, \omega_i;\, t_2, \omega_j) = \sum_{k=1}^{n} \mathbf{P}(t_0, \omega_i;\, t_1, \omega_k)\, \mathbf{P}(t_1, \omega_k;\, t_2, \omega_j). \tag{33}$$

When the phase space is a continuous manifold, the most typical case is that a *probability density* $p(t_0, \omega_0;\, t, \omega)$ exists for passing from the state ω_0 to the state ω in the interval of time (t_0, t). In this case the probability of passing from the state ω_0 to any of the states ω belonging to a domain G in the phase space Ω is written in the form

$$\mathbf{P}(t_0, \omega_0;\, t, G) = \int_G p(t_0, \omega_0;\, t, \omega)\, d\omega, \tag{34}$$

where $d\omega$ is an element of volume in the phase space.[16] For the probability density $p(t_0, \omega_0;\, t, \omega)$, the basic equation (33) takes the form

$$p(t_0, \omega_0;\, t_2, \omega_2) = \int_\Omega p(t_0, \omega_0;\, t_1, \omega)\, p(t_1, \omega;\, t_2, \omega_2)\, d\omega. \tag{35}$$

Equation (35) is usually difficult to solve, but under known restrictions we may deduce from it certain partial differential equations that are easy to investigate. Some of these equations were derived from nonrigorous physical considerations by the physicists Fokker and Planck. In its complete form this theory of so-called stochastic differential equations was constructed by Soviet authors S. N. Bernšteĭn, A. N. Kolmogorov, I. G. Petrovskii, A. Ya. Hinčin, and others.

We will not give these equations here.

The method of stochastic differential equations allows us, for example, to solve without difficulty the problem of the motion in still air of a very small body, for which the mean velocity c of its fall is significantly less than the velocity of the "Brownian motion" arising from the fact, because of the smallness of the particle, its collisions with the molecules of the air are not in perfect balance on its various sides.

Let c be the mean velocity of fall, and D be the so-called coefficient of diffusion. If we assume that a particle does not remain on the surface of the earth ($z = 0$) but is "reflected," i.e., under the influence of the Brownian forces it is again sent up into the atmosphere, and if we also assume that at the instant t_0 the particle is at height z_0, then the probability density $p(t_0, z_0;\, t, z)$ of its being at height z at the instant t is expressed by the formula

$$\begin{aligned} p(t_0, z_0;\, t, z) = {} & \frac{1}{2\sqrt{\pi D(t-t_0)}} \left[e^{-(z-z_0)^2/[4D(t-t_0)]} + e^{-(z+z_0)^2/[4D(t-t_0)]} \right] \\ & \times e^{-c(z-z_0)/(2D) - c^2(t-t_0)/(4D)} \\ & + \frac{c}{D\sqrt{5}}\, e^{-cz/D} \int_{(z+z_0-c(t-t_0))/(2\sqrt{D(t-t_0)})}^{\infty} e^{-z^2}\, dz. \end{aligned}$$

In Figure 4 we illustrate how the curves $p(t_0, z_0;\, t, z)$ may change for a sequence of instants t.

[16] Properly speaking, equation (34) serves to define the probability density. The quantity $p\, d\omega$ is equal (up to an infinitesimal of higher order) to the probability of passing in the time from t_0 to t from the state ω_0 to the element of volume $d\omega$.

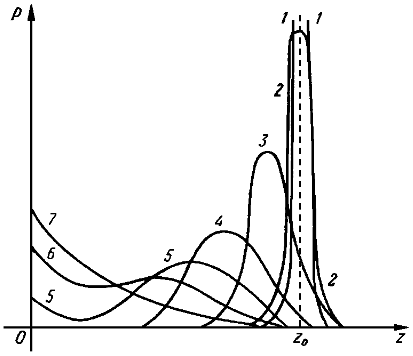

FIG. 4.

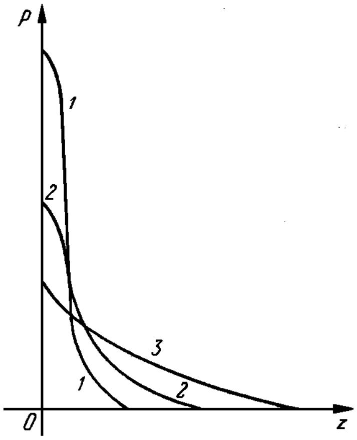

FIG. 5.

We see that in the mean the height of the particle increases, and its position is more and more indefinite, more "random." The most interesting aspect of the situation is that for any t_0 and z_0 and for $t \to \infty$

$$p(t_0, z_0;\, t, z) \longrightarrow \frac{c}{D}\, e^{-cz/D}; \tag{36}$$

i.e., there exists a limit distribution for the height of the particle, and the mathematical expectation for this height with increasing t tends to a positive limit

$$z^* = \frac{c}{D} \int_0^\infty z e^{-cz/D} dz = \frac{D}{c}\,. \tag{37}$$

So in spite of the fact that as long as our particle is above the surface of the earth, it will always tend to fall because of the force of gravity, nevertheless, as this process (wandering in the atmosphere) continues, the particle will be found on the average at a definite positive height. If we take the initial z_0 smaller than z^*, it will turn out that in a sufficiently great interval of time the mean position of the particle will be higher than its initial position, as is shown in Figure 5, where $z_0 - 0$.

For individual particles the mean values z^* under discussion here are only mathematical expectations, but from the law of large numbers it follows that for a large number of particles they will actually be realized: The density of the distribution in height of such particles will follow from the indicated laws, and, in particular, after a sufficient interval of time this density will become stable in accordance with formula (36).

What has been said so far is immediately applicable only to gases, to smoke and the like, which occur in the air in small concentrations, since the quantities c and D were assumed to be defined by a preassigned state of the atmosphere. However with certain complications, the theory is applicable to the mutual diffusion of the gases that compose the atmosphere, and to the distribution in height of their densities arising from this mutual diffusion.

The quotient c/D increases with the size of the particles, so that the character of the motion changes from diffusion to regular fall in accordance with the laws considered in section 5. The theory allows us to trace all transitions between purely diffusive motion and such laws of fall.

The problem of motion of particles suspended in a turbulent atmosphere is more difficult, but in principle it may be handled by similar probabilistic methods.

Suggested Reading

H. Cramér, *The Elements of Probability Theory and Some of Its Applications*, Wiley, New York, 1955.

W. K. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., Wiley, New York, 1957.

B. V. Gnedenko and A. Ya. Khinchin, *An Elementary Introduction to the Theory of Probability*, W. H. Freeman and Co., San Francisco, CA, 1961.

M. Kac, *Statistical Independence in Probability*, *Analysis and Number Theory*, Wiley, New York, 1959.

J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Van Nostrand, New York, 1960.

A. N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea, New York, 1950.

M. Loève, *Probability Theory*: *Foundations*, *Random Sequences*, Van Nostrand, New York, 1955.

E. Parzen, *Modern Probability Theory and Its Applications*, Wiley, New York, 1960.

W. A. Whitworth, *Choice and Chance*, Stechert, New York, 1942.

Review of the Paper "The Theory of Probability"

I consider the paper "The Theory of Probability" very well written in all respects. Written on a high scientific level, without condescending to the unqualified reader, it is nevertheless accessible to a wide range of readers interested in mathematics. However, the most valuable aspect of the paper is the exposition of relationships between general theory and practical applications, which has never been achieved so far in the principal presentations of probability theory; the author succeeds in demonstrating how the most abstract parts of the theoretical constructions are closely linked to specific practical demands. Therefore his presentation exhibits the theory of probability simultaneously as a well-balanced, completed logical construction and a powerful tool of natural science and technology, with these two images penetrating deeply into each other so that they make an organically connected whole.

10.X.1951. *A. Ya. Khinchin, Corresponding Member of the Acad. Sci. of the USSR*

From a Letter of A. D. Aleksandrov to A. N. Kolmogorov

Dear Andrey Nikolaevich!

Since I am responsible for completing the work on the monograph "Mathematics, Its Content, Methods, and Meaning," I turn to you regarding your paper.

The paper in its present form is difficult, and its appearance may create an impression that the foundations of probability theory are something uncomprehensible to a "simple mortal."

I read your article "Probability" in the *Great Soviet Encyclopedia*. I was very impressed with it. Of course, I am not an expert and I dealt with probability only in connection with physics, but I dare to express my opinion that I have never encountered anything comparable with regard to simplicity combined with depth of exposition. Working on the foundations of quantum mechanics, I came to the same concepts but could not give them such a deep and precise formulation. Your statement of the connection between the phenomenon and conditions is of great importance for quantum mechanics, for overcoming its purely statistical interpretation, which corresponds to Mises' concept of probability. (By the way, V. A. Fock told me that he views your concept of probability as extremely important.)

In order to convey your deep and important, and at the same time simple, ideas to the broadest possible readership, I would ask you to rearrange your paper. I suggest that, after mentioning the objective nature of statistical laws and criticizing briefly subjectivism, you immediately proceed to stating the connection between the phenomenon and conditions and presenting your concept of probability, in order not to obscure it with a deterministic scheme or any calculations. Your paper combining depth and simplicity of presentation will be the best part of our monograph. I urge you to simplify the presentation to make your ideas as accessible as possible. I would advise you to test its accessibility on some average reader having only a vague notion of the subject.

One more comment: When speaking of the relation of probability theory to measure theory, it is natural to point out that the distinction does not reduce to the fact that $\mathbf{P}(U) = 1$, but consists in the specific probabilistic setting that manifests itself, in particular, in the notion of independent events, which is apparently alien to the usual concepts in the study of areas and volumes.

15.II.1952 With best regards,

A. D. Aleksandrov

From a Letter of A. N. Kolmogorov to A. D. Aleksandrov

Dear Aleksander Danilovich!

Thank you for your letter regarding my paper for the monograph on mathematics. I certainly can fulfill your following requests:

1) To give examples explaining the notions of random variable, expectation, and variance when they are introduced, and to explain the Markov scheme by a particularly simple example.

2) To expose more completely the specific features of probability theory when speaking about the possibility of its formal inclusion into measure theory (I wrote about this in my "Fundamental Concepts of Probability Theory"; so I should have written something of this kind in the popular paper, but I simply forgot to do that).

The main difficulty in the exposition of the philosophical aspects of probability theory begins with clear presentation of the two dialectically combined statements:

1) There exists objective randomness.

2) There is nothing absolutely random.

I realize that to stand up for the right of existence of probability theory (and you know that this issue is not quite pointless) the former statement is most essential. But I think we have to speak of both sides even in a popular paper.

Anyway, I will bear in mind that in your opinion the paper does not highlight well enough the objective nature of probabilistic statements.

Once you have looked into my article "Probability" in the *Great Soviet Encyclopedia*, I would be interested to know your opinion about my article "The Law of Large Numbers," where I wrote only the general (not the social-economic) part. This question is of some interest because many of our economists strongly objected against my general part too. But the only point where I can admit my failure is that I called the LLN a "principle," which I perhaps should not have done.

26.II.1952

Yours sincerely,
A. N. Kolmogorov