METHODOLOGICAL NOTES

Is probability a "normal" physical quantity?

Yu. I. Alimov and Yu. A. Kravtsov

Joint-Stock Company TsITRON, Ekaterinburg; Small Enterprise GROT; and, Institute of General Physics, Russian Academy of Sciences, Moscow

(Submitted 6 March 1992)
Usp. Fiz. Nauk **162**, 149–182 (July 1992)

The informal aspects, arising in the interpretation of physical experiments, of the theory of probability and mathematical statistics are discussed. The conditions that verifying experiments must satisfy are presented and the role of heuristic (extralogical) assertions is analyzed using the example of mathematical expectation. The principal hypotheses implicit in experiments are enumerated: the principle of reproducibility ("the past will be repeated in the future"); the principle of reasonable sufficiency; and, the statistical principle ("better to predict something rather than nothing"). Considerable attention is devoted to Fisher and multisample confidence intervals. It is noted that Fisher confidence intervals are inconsistent. The arguments for introducing contrivances into practical calculations of probabilities are enumerated: incompleteness of any system of hypotheses; subjective estimates of probabilities; adjoining of statistical ensembles; nonstationariness and instability; rare phenomena; and, the use of classical probabilities and the law of large numbers. It is concluded that the relative frequency of appearance (empirical probability) is a "normal" physical quantity in the sense that it admits physical measurement. Its "abnormality" is manifested in the fact that it is burdened, more than other physical quantities, with conventions and hypotheses which must be specially checked (verified).

1. INTRODUCTION. INFORMAL ASPECTS OF PROBABILITY

To physicists probability is both a physical and mathematical quantity. But while as a mathematical concept probability appears to them to be sound and incontestable, there is often a sense of discomfort and reticence about probability as a physical quantity. It is in this connection that the question stated in the title arises: is probability a real, "normal" physical quantity, i.e., a quantity which admits physical measurements, or is it an unmeasurable, fictitious quantity?

If probability is in fact a normal physical quantity, then why are the results of statistical analysis of measurements constantly questioned? If, on the other hand, probability is somehow different from ordinary physical quantities, then how is its "abnormality" manifested?

Such questions became an integral part of science more than 200 years ago. In each age answers to them were sought in accordance with the ideas espoused then about the subject. Almost all the great physicists and mathematicians have had something to say, in one form or another, about the physical meaning of probability.

And yet the question still remains open. Why is it so difficult to come to an agreement about the physical meaning of probability and the accuracy with which it is measured? Maybe the impediment is something significant but difficult to grasp, something always remaining behind the scene?

We are inclined to believe that the significant element hidden in the physical interpretation of probability is a system of difficult-to-formalize hypotheses, agreements, and formal constructs which seem naturally, traditionally attached to the formal apparatus of the theory of probability, but which in reality are independent hypotheses that must be verified.

We discovered, to our surprise, that probability-physical constructs contain many more heuristic elements than one would expect, even given our many years of experience in working in mathematical statistics.

In the present paper we wish to discuss specifically the informal aspects of the concept of probability which arise in a physical context. We analyze, using the example of defining the mathematical expectation, the role of heuristic, extralogical assertions which are made in this seemingly well-assimilated field of knowledge. Essentially, we discuss the inconsistency of Fisher's procedure for analyzing the results of measurements, which has taken root in the practice of natural science in spite of doubtful formal constructs, and we describe an alternative procedure for finding multisample confidence intervals on the basis of simple and intuitively acceptable assumptions (see Sec. 5).

In addition, in Sec. 6 we discuss an extensive list of reasons for resorting to formal constructs in statistics, and in Sec. 7 we discuss probability from the standpoint of the concept of partial determinateness (predictability), which we feel is in better agreement with the logic of a physical experiment than the concept of algorithmic complexity.

2. VERIFYING EXPERIMENT

2.1. Mathematical "millstones" and heuristic "grist" of natural-science theories

In discussing probability as a physical quantity, i.e., a quantity that can be measured, it is helpful to recall the role which mathematical theory as a whole and the theory of probability in particular play in natural science. The theory

 0038-5670/92/070606-17$03.00 © 1992 American Institute of Physics

of probability functions in natural science in the same manner as other mathematical disciplines.

Mathematics is a rich special language, whose grammar has a powerful, formal, built-in logic. Natural scientists have often noted the "incomprehensible effectiveness of mathematics" and its indispensability as a unique linguistic tool (the best known paper on this subject is probably that of E. Wigner[1]).

At the same time, of course, practical work does not reduce merely to manipulation of words. Correspondingly, natural-science theories, being ultimately aimed at practical applications, do not reduce to mathematical constructs. The limited role of mathematics in natural science was described well by T. Huxley: "Like millstones, mathematics grinds anything that is poured into it, and just as you will not obtain wheat flour by grinding weeds, you will not obtain the truth from false assumptions by writing out entire pages of formulas" (Ref. 2, p. 106).

The fact that mathematical millstones require heuristic grist from without is due to the nature of formal-logical inferences. Such inferences—including mathematical calculations—always take the implicative form "if ..., then ... ." The official duties of mathematics in applications reduce, in general, to the solution of the following problem: To calculate an output quantity $Y = F(X)$ from a given initial value X and a given (possibly, in an implicit form) mapping F. The mapping F represents the millstones—the mathematical model of reality. Elaborating Huxley's words, we underscore the fact that the unavoidable heuristic grist or, as stated more dryly in Ref. 3, the *extralogical component* includes the choice of not only the working value of the starting quantity, but also the mathematical millstones. With poor millstones one cannot obtain good flour even from wheat.

In applications, mathematics also plays an unofficial role, providing natural scientists the language itself for describing the quantities X, Y, and the mapping F. Although mathematics has no direct responsibility here, this purely linguistic role of mathematics is no less important than the above-mentioned official computational role.

The choice of the working value of the starting quantity X turns out to be essentially heuristic, even when this value is obtained by means of measurements. After all, measurements and experiments in general are not completely formalized operations for the simple reason that they concern reality and not some formal system. But then, the investigator has the right to postulate the starting working value of X without resorting to measurements of this quantity. With regard to the output quantities Y, however, a more stringent tradition has evolved in natural science: The computed value of Y must always be compared with reality with the help of measurements. Making such a comparison comprises the experimental check of the adequacy—the *verification*—of the entire theory, including also the choice of the working values of X. Otherwise, the theory will not be a theory of natural science, but only a mathematical construct.

2.2. Metrological precepts of a verifying experiment

In the final results of a verifying experiment—figuratively speaking, the decision of a court of last instance—everything must be intuitively clear, immediately convincing, at least for most specialists in a given field of research. For this, it is first necessary to adhere to the following fundamental metrological precepts:

1°. A model of the object of measurement must be carefully constructed

Otherwise it will be unclear whether or not the mathematical quantity Y refers to the physical quantity measured in the verifying experiment. The construction of a model of the object of measurement is an organic component part of verification. We now discuss this in detail.

In performing any measurements the experimenter mentally replaces the real object of measurement by a model. The measured quantities are, strictly speaking, parameters of this model and not of the real object. The true value of the measured quantity is the value of the model parameter that would be obtained as the result of an indisputable experiment performed on an ideal object, which is the exact materialization of the model. However, there can always be some qualitative disparity between the properties of a real object and that which is taken as the measured quantity. This disparity engenders the "error of disparity between the model and the object" (Ref. 4, p. 14), included in the methodological and sometimes also in the systematic errors. If the error of disparity is too large, then a new model is constructed.

An adequate model of the object of measurement must guarantee that the concept of the true value of the measured quantity is itself sensible. For example, suppose that the size of a spherical body is being measured, say, small pellets or the earth. Let the model be a sphere. Correspondingly, the problem is to estimate a unique value of the diameter of the body. Suppose, however, that measurements of several diameters of the spherical body gave results which differ by amounts greater than the error of the means chosen to perform the measurements. Then it must be stated that in this specific situation the model in the form of a sphere is not adequate to the object. Consequently, the problem of estimating the true unique value of the diameter of the body, as posed above, is meaningless: There is no such value within the existing accuracy of the measurements. Such a negative result, obtained in constructing a model of the object of measurement, in itself can indicate that the entire theory being verified is inadequate.

We give two more examples. Consider an electric signal generator, whose structure is unknown to us. We shall try to determine the structure of the generator from the form of the generated pulses. If the pulse shapes are reminiscent, for example, of the process of charging and discharging of a capacitor, then it is natural to take a relaxation oscillator as a primary model. Such a model may turn out to be satisfactory with a rough fit of the circuit parameters (capacitance C and resistance R) to the shape of the generated pulses. If it is found later that the difference between the real and model pulses is greater than the error of the measuring instrument, then the model of a simple RC oscillator must be acknowledged to be inadequate to the observed process, so that a more complicated model will be required.

This example illustrates a general characteristic feature of inverse problems: As data accumulate, first, the parameters of the simple model are made more precise; there then comes a time when the model as a whole is radically reconstructed.

For the second example, consider the measurement of

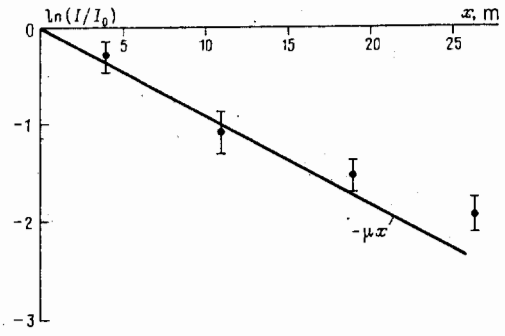

FIG. 1. Illustration for the question of choosing a linear model $-\mu x$ for the experimental dependence $\ln[I(x)/I_0]$. Too large a spread of the experimental points indicates that the linear model is unsuitable and the idea of a constant absorption coefficient μ is inapplicable.

the attenuation coefficient of a wave field in an absorbing or scattering medium. Suppose that in the experiment the intensity I is measured as a function of the distance x; this dependence is presented in a logarithmic scale in Fig. 1. Assuming that the intensity varies exponentially as $I(x) = I_0 \exp(-\mu x)$ (Bouguer's law), the attenuation coefficient μ can be found by fitting the linear function $-\mu x$ to the experimental points (i.e., the logarithms of the measured values).

If the difference between the experimental values $\ln(I/I_0)$ and the linear function $-\mu x$ is greater than the measurement error, then the interpretation of μ as the attenuation coefficient of a uniform medium becomes meaningless. In this case the model of the medium must itself be reexamined and/or the experiment must be repeated in greater detail or with higher accuracy (incidentally, the accuracy may be limited not only by physical factors but also by cost factors).

The development of a model of the object of measurement is, on the whole, an informal operation which requires deep knowledge about the subject. Calculations sometimes are so tightly intertwined with the experimental investigations that one can talk about a single *theoretical-experimental process of construction of a model of the object.*

Suppose, further, that an adequate model of the object of measurement has been constructed and is utilized in a verifying experiment. Now, in order for the results of this experiment to be intuitively clear and convincing, the final quantitative indicators of adequacy of the theory must be simple and not associated with any complicated formal constructs. We are talking about indicators which characterize the accuracy of the final estimate adopted for the true value of the quantity Y. After all, the matter has now reached the point of empirical substantiation of the validity of the theory!

Contrary to scholasticism, in natural science the basic pragmatic thesis was adopted immediately: *The validity of a theory is ultimately based on experiment and not on some other theory* (say, on the works of Aristotle or K. Marx).

Suppose that for numerical analysis of the final results of a verifying experiment we nonetheless construct a quite complicated mathematical theory. Such a secondary theory will engender, correspondingly, complicated quantitative indicators of adequacy of the initial theory. These indicators pertain to reality, evidently, in the same measure as the secondary theory is adequate as a whole. We have arrived at logical cycling: In the process of verifying the primary theory we have found a more difficult and, more importantly, hopeless problem of verifying a secondary theory. This new problem is more difficult than the initial problem precisely when the secondary theory is formulated mathematically and is quite complicated. The hopelessness is due to the logical cycle: If we start to verify, as before, also the secondary theory, then we shall need to verify the difficult, formalized, "third generation" theory which arises, i.e., the theory of the indicators of accuracy of the secondary theory. It is obvious that by truncating the chain we risk casting doubt on even the primary theory.

In short, in order to avoid the logical cycling the model used to compare the final results of a verifying experiment with the computed output results of a theory of natural science must be sufficiently simple and informal. The model used for making the comparison, by its very nature, is the heuristic grist, an extralogical component, and it is not the mathematical millstones.

We can state that along the path from theoretical work to verification the language employed is unavoidably reduced to maximally deformalized "language of observation." Otherwise, the verification process cannot give intuitively and informally convincing results. Traditionally, in natural science it is desirable that together with the requirement 1° two additional requirements be satisfied:

2°. It must be demonstrated convincingly that the corresponding experimental result is reproducible

3°. The systematic errors in the experimental result being reproduced must be evaluated convincingly

It is sometimes necessary to evaluate, with the highest possible accuracy, the true values of the input or some intermediate quantities of the theory (and not only of the output Y). This happens, for example, if one wants to verify or, as it is often said, identify the model of the object, i.e., the millstones, itself. In such situations it is also necessary to adhere to the fundamental metrological precepts 1°–3° and the attendant requirement that the final indicators of the accuracy of measurement be simple and informal.

Thus mathematics and metrology both serve as indispensable supports for the entire quantitative natural science. The requirements 1°–3° are necessary when performing any measurements, including measurements of probabilities or, more generally, mathematical expectations.

We are not belaboring the obvious. The theory of statistical analysis of data, called mathematical statistics and associated with R. Fisher, has strongly influenced the extant metrological standards.[5] This theory is not in complete agreement with the traditional metrological precepts (see, for example, Refs. 12–18 for a critique of Fisher's statistics). Thus, for verification Fisher proposes that the confidence intervals be calculated on the basis of a by no means simple mathematical model of a verifying experiment.

The ideology of Fisher's mathematical statistics is analyzed below in Sec. 5. Separate remarks concerning Fisher's statistics are made in the course of the exposition even earlier; they are addressed to readers who are familiar with Fisher's mathematical statistics.

2.3. With what is the theory of probability concerned

From the viewpoint of the naturalist, the theory of probability is concerned with the same thing as any other branch of mathematics: It invents and provides users with the mathematical millstones for logical grinding of the heuristic grist.

The character of the grist, the millstones, and the flour in the domain of the theory of probability is somewhat peculiar to this theory. The basic probability-theoretic concepts employed in applications are mathematical expectation, probability, and probability distribution. We refer to them collectively as *probability characteristics* (of random quantities or random events).

On the basis of the statistical interpretation, probability characteristics pertain to an imaginary object. In physics this object is termed an ensemble, while in mathematical statistics and the theory of probability it is referred to as the universe of objects or trials (observations).

The peculiarity of the theory of probability is that usually the grist fed into its mathematical millstones and the output are considered to be the probability characteristics. In other words, the theory of probability converts one set of probability characteristics into another.

We illustrate this for the very simple example of a probability-theoretic relation for the sum of random quantities $Y = \Sigma a_i X_i$:

$$M[Y] = M[\sum_{i=1}^{m} a_i X_i] = \sum_{i=1}^{m} a_i M[X_i]; \qquad (2.1)$$

here M is the mathematical expectation operator, X_i are random quantities, and a_i are constants. In Eq. (2.1) the given quantities $M[X_i]$,..., $M[X_m]$ and the model parameters a_1,..., a_m are the grist; the output is the computed value of the mathematical expectation of the linear function $Y = \Sigma a_i X_i$.

In order to determine how probability-theoretic relations arise and the meaning of such relations, it is necessary to appeal to physical analysis of the basic concepts. Sections 3 and 4 are devoted to these questions, which are the central questions in this paper. The theory of probability, as any mathematical discipline, obviously cannot claim to calculate the accuracy and reliability of the grist and the millstones (i.e., the accuracy of the initial quantities) of the model employed and of the final results of a verifying experiment. It is precisely such a claim that can be seen in the manner in which Fisher's mathematical statistics approaches the problem of verification.

3. HYPOTHESES IMPLICIT IN EXPERIMENTS

3.1. Principle of repeated reproducibility: "the past will be repeated in the future"

We now return to the metrological precepts 1°–3° and we examine their most important feature: These precepts ensure, no more and no less, that the natural sciences solve their main problem—to predict the results of impending experiments.

Quantitative predictions in natural science take the form of the following assertion: the realization of definite controllable conditions U in an experiment will always result in the same (within the limits of the stipulated accuracy) measured result Y. We designate this prediction by the symbol $U \to Y$ (from U to Y). To verify a theory means to verify the prediction made by it.

The description and realization of the conditions U as well as the measurement of the predicted result Y are performed in accordance with the precepts 1° and 3°, which direct the investigator to construct carefully a model of the object of measurement, taking all measures to eliminate systematic errors and to evaluate those errors that cannot be eliminated. If these requirements are not satisfied, then one cannot speak at all about verification of the prediction $U \to Y$.

According to the condition 2°, in order to verify a prediction the conditions U must be reproduced many times and, correspondingly, the measurement Y must be repeated many times. In natural science, traditionally, a predicted result is seriously accepted only if it is reproduced in a quite long series of uniform trials without any mysterious deviations. All experience in natural science teaches that repetition of trials is a necessary though not absolute antidote to random errors, gross blunders, and juggling of facts.

Of course, verification of a prediction even in accordance with the traditional requirement of reproducibility is not infallible. Such a verification is itself a prediction associated with the concept of empirical induction "it has happened many times in succession and therefore it will also happen in the future" or more briefly "tomorrow will be the same as today." It is well known—at least from everyday life—that this empirical-inductive prediction is by no means infallible. In particular, this principle could pose a well-known danger in the analysis of unique, expensive, unplanned, and some other experiments, say, experiments in space, deep in the ocean, and under other difficult conditions. This principle of "tomorrow will be the same as today" is conservative and is organically incapable of revealing (it is not "tuned to") new trends and changes.

In spite of this, science does not know of any methods of verification which are more convincing and reliable than multiple repetition of experiments satisfying the conditions 1°–3°. However, the natural sciences do not claim to be infallible.

3.2. Principle of reasonable sufficiency

When a series of uniform trials is performed, there arises the question of what the minimum length Q of this series should be in order for reproducibility of the experimental results to be convincing. It is important to realize that this question concerns the heuristic grist and thus it cannot be resolved by formal-logical methods. Apparently, from time to time a need is felt to recall the fact that heuristic components are unavoidable in investigations in natural science. Thus E. L. Feinberg (Ref. 3, p. 35) stated concerning extralogical judgment regarding the sufficiency of a given experimental check: "... It occurs in any, even completely ordinary experiment, which remains limited, regardless of how it is modified or expanded or how many times it has been repeated. There still comes the time when the investigator must say: 'Enough, I am now *convinced* that such and such a law is correct.' This "I am convinced" expresses an extralogical judgment in science which is unavoidable and which is part of the base of the process of gathering knowledge."

One can only guess the moment when a conjecture or

hypothesis becomes certain. Most likely, an internal criterion, such as *reasonable sufficiency*, operates. Judgments accepted by the community of specialists are accepted as significant, so that reasonable sufficiency is oriented, directly or indirectly, toward the reigning scientific paradigm.

The choice of the acceptable dispersion ΔY of the measurements of the quantity Y in a series of trials and the estimation of the systematic errors also has to be made, essentially, extralogically. In short, all of this is done according to the closest precedents. If there are many precedents and they are uniform, sound empirical induction is in evidence.

3.3. Statistical prediction: "better to predict something rather than nothing"

Measurement of probability characteristics is especially closely associated with prediction. We shall trace this relation in Sec. 4 using the example of mathematical expectation. The overall situation here is as follows: Measurement of probability characteristics of a quantity Y must start with the measurement of the corresponding *statistical characteristics* S_Y, which are also often called *sample statistics*; in physics they are called *empirical characteristics*. Thus the probability characteristics are a kind of superstructure built on the empirical data—statistical (sample) characteristics.

From the standpoint of prediction, the logic here is completely understandable. The investigator, having become convinced that the quantity Y cannot be predicted accurately enough, arrives from the unsuccessful "dynamical" prediction $U \to Y$ to a rougher statistical prediction $U \to S_Y$, hoping that the latter prediction will be *successful* and to some extent *helpful* besides. In other words, here the pragmatic idea "*better to predict something rather than nothing*" is followed.

The statistical characteristics of quantities which cannot be predicted in a dynamical sense are, in reality, often reproducible, i.e., they remain almost constant from one sufficiently large sample to another, which is why they can be predicted quantitatively. This version of reproducibility of an experimental result is called the *statistical stability* of the result.

The verification of a prediction of a statistical characteristic should also be made, ideally, in accordance with the requirement that an experimental result should be reproducible in a series of uniform trails. There are no grounds for excepting statistical predictions from this traditional requirement in natural science. The only peculiarity here is that when a statistical prediction $U \to S_Y$ is verified, each trial is a component part and it is secondary compared with the initial trials in which the values of Y were measured. It consists of calculating the statistical characteristic S_Y for a *series* of initial trails. In short, multiple repetition of secondary trials means obtaining and analyzing many samples. We term such verification *multisample* verification. We note that Fisher's mathematical statistics is oriented specifically toward multisample verification of statistical predictions.

There also appear to be no grounds for making other exceptions for statistical predictions $U \to S_Y$ as compared with the predictions $U \to Y$. Everything we have said above about verification of natural-science theories and their predictions must also apply to cases when these predictions are statistical. Here the final results of a verifying experiment must also be intuitively clear, which requires that these results be described in a simple language and that precedents be used as the final argument. In particular, here the choice of the lower limit $Q_{\min}$ for the required length of the series of checking trials and the choice of the acceptable dispersion ΔS_Y in the results of measurements of S_Y are also, essentially, extralogical. Fisher's mathematical statistics disregards even these important circumstances.

4. MATHEMATICAL EXPECTATION AND PROBABILITY AS PHYSICAL QUANTITIES

4.1. Sample average

The measurement of the mathematical expectation $M[Y]$ of a random quantity Y starts with the measurement of the statistical (sample, empirical) average

$$M_n[Y] \overset{\text{not}}{=} \frac{1}{n}\sum_{s=1}^{n} Y(s); \tag{4.1}$$

here, for clarity of exposition, we have introduced the symbol not—"equal up to the notation"; s is the trial number and $Y(s)$ is the value of the quantity Y measured in the sth trial. The sequence

$$\{Y(1), \ldots, Y(n)\} \overset{\text{not}}{=} \{Y(s)\}_1^n \tag{4.2}$$

is said to be a sample of size n.

Let the quantity $Y(s)$ be represented by an indicator $I_A(s)$ for an event A, defined as

$$I_A(s) = \begin{cases} 0 \\ 1 \end{cases}, \text{ if in the } s\text{th trial event } A \begin{cases} \text{did not occur} \\ \text{did occur} \end{cases}. \tag{4.3}$$

Then the quantity

$$\omega_n(A) \overset{\text{not}}{=} M_n[I_A] = \frac{1}{n}\sum_{s=1}^{n} I_A(s) = \frac{n(A)}{n}, \tag{4.4}$$

which is the sample average of the indicator I_A, is the relative frequency of appearance of the event A in the sample $\{I_A(s)\}_1^n$. This quantity is equal to the ratio of the number of trials $n(A)$ in which the event A occurred to the total number of trials n. Everything said about $M_n[Y]$ is also applicable to the special case, under consideration, of the relative frequency ω_n.

The quantity $M_n[Y]$ is one of the statistical characteristics of the quantity Y, so that according to what was said in Sec. 3.3 about the statistical characteristics, we must evaluate experimentally the reproducibility of sample averages. To do this we must obtain, ideally, under the conditions of the experiment U a sufficiently large number Q of equivalent samples $\{Y_{(k)}(s)\}_1^n$, $k = 1,\ldots,Q$ of size n. Here $Y_{(k)}(s)$ is the measured value of Y in the sth primary trial in the kth sample, so that we enumerate the secondary trials with the index k.

Equality of the sizes of all these samples is a necessary condition for the secondary trials to be uniform: We divide the entire available primary sample $\{Y(s)\}_1^N$ into Q equivalent subsamples, so that the total number of all measurements (trials) is equal to $N = Qn$. The number of subsamples Q should not be less than some threshold value $Q_{\min}$, representing the lower limit of the number of secondary trials: $Q > Q_{\min}$. The value of $Q_{\min}$ is determined by extralogical considerations and is adopted according to precedents.

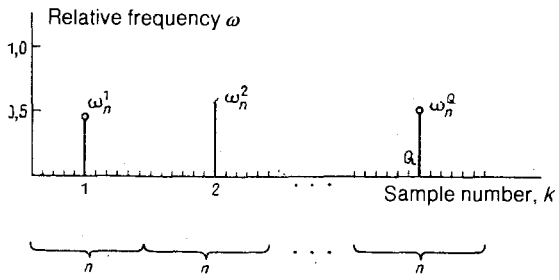

FIG. 2. The relative frequency changes from one sample to another. As the size n of a sample increases the spread in the quantities $\omega_n^k, k = 1,...,Q$, decreases.

The k th secondary trial is completed by calculating the subsample average according to the algorithm (4.1):

$$M_{k,n}[Y] \overset{\text{not}}{=} \frac{1}{n}\sum_{s=1}^{n} Y_{(k)}(s), \quad k = 1,...,Q. \tag{4.5}$$

In the case when the measured quantity Y is the indicator I_A of the event A, Q series of trials will yield Q values of the relative frequencies of appearance $\omega_n^1,...,\omega_n^Q$ (Fig. 2). These values vary from series to series, as is characteristic of physical measurements.

Thus the relative frequency ω_n^k behaves as an ordinary physical quantity, i.e., a measurable quantity. The "normal" physical properties of the relative frequency become less obvious as soon as we attempt to interpret ω_n in terms of probability. We return to this question in Sec. 4.5.

4.2. Decrease of the dispersion with increasing sample size

In the scheme under consideration, the primary unpredictable quantity $Y(s)$ is replaced by the secondary quantity $M_{k,n}[Y]$, which, generally speaking, is also unpredictable. The point of making this replacement is that the secondary quantity still turns out to be more predictable, since its dispersion for $k = 1,...,Q$ and $n \gg 1$ is usually much smaller than that of $Y(s)$ for $s = 1,..,N$.

It is useful to describe the degree of dispersion in the secondary sample

$$\{M_{k,n}[Y]\}_1^Q \overset{\text{not}}{=} \{M_{1,n}[Y], ..., M_{Q,n}[Y]\} \tag{4.6}$$

with the help of some simple quantitative characteristic. The dispersion of the set of numbers (4.6) is most simply characterized by the interval $[\overline{Y}_{\min}, \overline{Y}_{\max}]$ which contains all values of the quantity $M_{k,n}[Y]$ in the sample (4.6):

$$\overline{Y}_{\min} = \min_{k=1,Q} M_{k,n}[Y], \quad \overline{Y}_{\max} = \max_{k=1,Q} M_{k,n}[Y]. \tag{4.7}$$

Thus

$$M_{k,n}[Y] \in [\overline{Y}_{\min}, \overline{Y}_{\max}], \quad k = 1,...,Q. \tag{4.8}$$

The limiting values $\overline{Y}_{\min}$ and $\overline{Y}_{\max}$ depend on the number of trials n in a single sample and on the total number of samples Q.

Adhering to the position presented in Sec. 2.3, we underscore the fact that the dispersion of the subsample averages *must be characterized simply*, since we are talking now about an empirical estimate of the validity of the theory. The theory itself can be formalized and as complicated as desired. However, in verifying the theory it is no longer appropriate to utilize complicated formalisms.

TABLE I. Actual limits of variation of the quantity $M_{k,n}[Y]$ in a secondary sample for the die rolling problem.

Q	n	$N = Qn$	$\overline{Y}_{\min}$	$\overline{Y}_{\max}$	$\Delta\overline{Y}$
8	40	320	2,98	3,72	0,74
8	160	1280	3,31	3,60	0,29

For $n \gg 1$ the total range $\Delta\overline{Y} = \overline{Y}_{\max} - \overline{Y}_{\min}$ of the fluctuations of the subsample averages (4.1) is often indeed much smaller than the total range of the fluctuations of the primary quantity Y. As the number n of trials in the subsample increases, the quantity $\Delta\overline{Y}$ decreases appreciably, even if the number Q of subsamples is fixed, since the total size $N = Qn$ of the sample being analyzed increases.

An example of such a manifestation of stability of subsample averages is given in Table I, taken from a report (Ref. 14, Sec. 1.2) of laboratory work performed in a "non-Fisher" course in probability theory.[14-16] The primary test was rolling a standard die and recording the number of spots obtained $Y(s)$. One can see from Table I that the values of $M_n[Y]$ cluster around the classical average $\overline{Y}_{\text{class}} = (1/6)(1+2+3+4+5+6) = 3.5$, but as the sample size increases the dispersion relative to this value of $\overline{Y}_{\text{class}}$ decreases from $\Delta\overline{Y} = 0.74$ for $n = 40$ and $Q = 8$ to $\Delta\overline{Y} = 0.29$ for $n = 160$ and $Q = 8$.

4.3. Interval statistical prediction

We state at the outset that the relation (4.8) was satisfied repeatedly and without a single failure in the series of uniform secondary trials performed (to readers who are confused by our choice of interval on the right-hand side of the relation (4.8) we recommend that they imagine that this choice was made even before the start of the experiment). On the basis of the principle of empirical induction "it has happened many times in succession and therefore it will happen in the future" we suppose that the relation (4.8) will also hold in all secondary trials:

$$M_{k,n'}[Y] \in [\overline{Y}_{\min}, \overline{Y}_{\max}], \quad \forall\, k > Q \text{ and } \forall\, n' > n. \tag{4.9}$$

We enumerate the future subsamples by the numbers $k = Q+1, Q+2,...$.

The interval statistical prediction (4.9) is, essentially, the main result of the multisample empirical estimation (stability) of averages. This heuristic prediction was obtained on Q subsets of size n. But it was formulated for sample sizes $n' > n$ because the reproducibility of averages usually improves as the size of the subsamples increases (in particular, this improvement is demonstrated in Table I).

4.4. Point statistical prediction. Mathematical expectation

Let $\Delta\overline{Y}$ be sufficiently small, i.e., $\Delta\overline{Y}$ is less than the allowed dispersion of the subsample averages, which is chosen based on precedents. We call this favorable situation stability of averages or, more generally, statistical stability. In the present situation we replace the interval estimate-prediction (4.9) by a somewhat rougher but simpler point estimate

$$M_{k,n}[Y] \cong M[Y], \quad \forall\, k > Q \text{ and } \forall\, n' > n; \tag{4.10}$$

here $M[Y]$ is a number chosen from the interval

$[\overline{Y}_{\min}, \overline{Y}_{\max}]$ to represent this interval. From the standpoint of the experimenter, this number is what is termed the measured value of the mathematical expectation of the random quantity Y.

The rule for choosing the number $M[Y]$ must be simple, appealing only to common sense and precedents, but not to any formal models. The midpoint of the interval can be used to represent the interval. In our case, this means the choice $M[Y] = (\overline{Y}_{\min} + \overline{Y}_{\max})/2$. The rule $M[Y] = M_N[Y]$, where $N = Qn$ is the size of the entire experimental sample, is employed more often. There is no point in "philosophizing" when choosing the values of $M[Y]$: The error of measurement of $M[Y]$ must still be taken equal to approximately $\Delta\overline{Y}$.

4.5. Statistical (empirical) probability as mathematical expectation

The statistical (in the physical jargon, empirical) probability $p(A)$ of a random event A is the mathematical expectation of the indicator I_A introduced by the relation (4.3):

$$p(A) \overset{\text{not}}{=} M[I_A]. \tag{4.11}$$

In short, from the viewpoint of the experimenter, the statistical probability $p(A)$ is a particular case of mathematical expectation. This agrees with the relatively new trend of constructing a theory of probability not on the basis of a probability measure, as done in Kolmogorov's set-theoretic axiomatics, but rather on the basis of an averaging operation. This trend is most pronounced in P. Whittle's book of Ref. 19. To complete the picture, we indicate how the probability distribution of a random quantity is measured: The method actually reduces to measuring a finite set of probabilities, i.e., once again measurement of mathematical expectations.

In accordance with the properties of mathematical expectation, as the sample size increases (the trials U are assumed to be uniform) the dispersion of $p(A)$ often decreases. This agrees with R. von Mises's concept of frequency of appearance.[20] One can only agree with V. N. Tutubalin that "the conditions for practical applicability of the theory of probability are now interpreted according to R. von Mises" (Ref. 21, p. 143).

In general, von Mises's construction of probability theory does not fit within A. M. Kolmogorov's axiomatics.[22] A critique of this axiomatics is contained, in particular, in the publications of one of the present authors.[12,13,23,24]

Adopting von Mises's correspondence between the relative frequency of appearance and probability

$$\lim_{N\to\infty} \omega_N(A) = p(A), \tag{4.12}$$

we transfer, at the same time, to probability the explicit and implicit assumptions about the relative frequency. Not all of these assumptions fit naturally into Kolmogorov's set-theoretic interpretation,[22] which also contains a system of assumptions and conventions. It is sufficient to mention, for example, the convention about the existence of a statistical ensemble (universe), satisfying a definite probability measure, or the convention about random functions (the complexities and conditionalities of this concept are described, for example, by A. M. Yaglom in Ref. 25).

The coexistence of A. N. Kolmogorov's abstract set-theoretic approach and R. von Mises's frequency interpretation, oriented toward experiment, also leads to the reverse process—transfer of the assumptions adopted in the set-theoretic constructs to the domain of practical statistics. In practice this means, for example, the formal construction of an esemble of values (universe) for a limited sample obtained in an experiment. It is obvious that this construction of an ensemble for a specific experiment opens up extensive possibilities for arbitrariness, if not for abuse, and by no means always leads to positive results.

4.6. Random and indeterminate quantities

Thus, the above-described multisample procedure for measuring the mathematical expectation $M[Y]$ contains organically an empirical estimation of the stability of averages $M_{k,n}[Y]$. In applied probability theory, only those unpredictable quantities whose subsample averages have been found experimentally to be stable are now called random. Consequently, the mathematical expectation exists (more accurately, it is assumed to exist) not for any unpredictable quantity, but only for a random quantity.

The result of empirical estimation of the stability of averages can also be negative: The interval in the prediction (4.9) is sometimes too large in order to be able to go over to point estimation (4.10). If, however, Q is small, then the

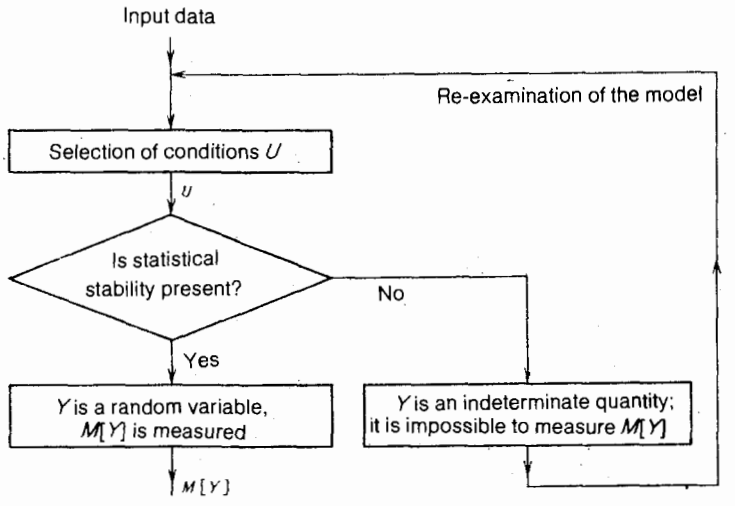

FIG. 3. Procedure for measuring the mathematical expectation $M[Y]$. In the procedure, measurements are rejected if Y is found to be an indeterminate quantity. In this case it is recommended that the model of the phenomenon be reexamined.

matter never reaches the point of adopting the prediction (4.9). Unpredictable quantities, whose subsample averages are not shown experimentally to be stable, are now customarily called *indeterminate quantities* (Ref. 21, pp. 6–7, 144 and 145; Ref. 26, p. 24). For such quantities there is simply no definite mathematical expectation; more accurately, its existence is doubtful. Here the situation is approximately the same as in the above-mentioned attempt to measure the single true value of the diameter of a body, whose shape, as becomes clear in the course of the experiment, differs appreciably from spherical.

The scheme of the branching procedure, characterized above, of measuring the mathematical expectation is presented in Fig. 3. The controllable experimental conditions, supplemented by a description of the method for constructing equivalent subsamples, are designated by U.

In short, many traditional concepts of the theory of probability and mathematical statistics, starting with the concepts of mathematical expectation and the universe, are not applicable to an indeterminate quantity. This is how V. N. Tutubalin expressed himself concerning this point (Ref. 27, p. 7): "All imaginable experiments can be divided into three groups. The first group consistes of good experiments in which the result of an experiment is always completely stable. The second group consists of experiments which are not as good, where there is no complete stability, but there is statistical stability. The third group consists of very poor experiments, when even statistical stability does not exist. In the first group everything is obvious without the theory of probability. In the third group the theory of probability is useless. The second group is the real domain of application of the theory of probability, but there is hardly any guarantee that the experiment of interest falls into the second and not the third group."

Agreeing essentially with this statement, we note that even in the third group of experiments (which concern indeterminate quantities) the statistical (sample, empirical) characteristics still can be quite useful, though the probability characteristics are inapplicable here. The statistical characteristics can also be classified as probability-theoretic concepts. Everything depends on the point of view as to with what the theory of probability is concerned and how it operates (see Sec. 2.3).

Statements of the type "if the trials are uniform, then statistical stability is present" are encountered in the literature. Such a statement makes sense, evidently, only if the *controllable experimental conditions* are uniform (i.e., constant). It is absurd to speak about constancy of all experimental conditions.

The facts indicate, however, that sometimes a number of significant circumstances fall outside the field of view of the experimenter. Then there is no stability, even statistical stability, though all controllable experimental conditions are constant.

This happens even in fundamental physics, where experiments as a whole are significantly cleaner than in other fields. In Rutherford's laboratory the results of an experiment with a new radioactive substance one day became really chaotic, though the conditions of the experiment did not change. The investigators did not immediately suspect that the new substance was a gas (now called radon). None of the previously discovered radioactive substances was gaseous. The experimental conditions were further refined: Drafts were eliminated and smoking near the experiment was stopped. After this, statistical stability was restored. This episode is described in D. Danin's book "Rutherford."

The experimenter is never completely guaranteed to be free of such surprises. In technology [for example, in reliability and quality control problems (Ref. 26, pp. 24–27; Ref. 16)] indeterminate quantities, unfortunately, are not rare. They are encountered even more often in economics and sociology. For this reason *N.* Wiener even excluded these fields from the domain of cybernetics as a highly mathematicized "interdisciplinary" discipline. In his last scientific memoir[28] he wrote: "The success of mathematical physics led the social scientists to be jealous of its power without quite understanding the intellectual attitudes that had contributed to this power... . Just as primitive peoples adopt the Western modes of denationalized clothing and of parliamentarism out of a vague feeling that these magic rites and vestments will at once put them abreast of modern culture and technique, so the economists have developed the habit of dressing up their rather imprecise ideas in the language of the infinitesimal calculus... . Difficult as it is to collect good physical data, it is far more difficult to collect long runs of economic or social data so that the whole of the run *shall have a uniform significance* [the italics are ours—authors]... . Under the circumstances, it is hopeless to give too precise a measurement to the quantities occurring in it. To assign what purports to be precise values to such essentially vague quantities is neither useful nor honest, and any pretense of applying precise formulae to these loosely defined quantities is a sham and a waste of time."

The example of the discovery of radon shows that *it is precisely the output characteristic of an experiment—the existence of statistical stability*—and not the input characteristic of the experiment—stability of the controllable conditions—that must be taken as the final criterion for uniformity of the trials. In other words, perhaps only inverse statements of the type "if statistical stability exists, then the trials are uniform" are completely acceptable. Such inverse statements do not claim anything significant. They merely introduce yet one more term—"uniformity of trials"—for the concept of statistical stability.

For all that, the experience of the natural sciences and their applications teaches that the main means of increasing the reproducibility of an experimental result is to ensure that as many of the experimental conditions as possible are stable. Otherwise, a contribution due to sloppiness will also be added to the unavoidable dispersion. Thus, on the subject of mathematization of quality control it is vividly stated in Ref. 27 (p. 5): "... in order to apply probability theory to production quality analysis it is necessary first to create a well-organized production process, in which there is no drunkenness, unexcused absences, rush work, unacceptable raw materials, worn-out technological equipment, etc. The theory of probability is something like butter in porridge: First it is necessary to have the porridge."

We note, in passing, that the concept of randomization of an experiment, characteristic for Fisher's mathematical statistics, does not agree with this traditional view of things.

It is a different matter that in applied investigations it is expedient to choose the controllable experimental conditions at the last stage of laboratory investigation to be the

conditions expected in future use of the product being developed.

4.7. Interim summary

We now summarize what we have said about sample averages and mathematical expectation, utilizing metrological categories. The calculation of the quantity $M_n[y]$ according to the simple formula (4.1) and according to the measured values $Y(s), s = 1,...,n$, of the primary quantity Y is a form of indirect measurement. The algorithm (4.1) for measuring $M_n[Y]$ is thus distinguished by precision and indisputability [if, of course, the values of $Y(s)$ have already been measured]: It is identical to the formal definition of the sample average. Here it is not necessary to develop a model of the object of measurement. The algorithm (4.1) is applicable to any quantities Y, irrespective of whether or not statistical stability exists.

Measurement of mathematical expectation $M[Y]$, being more indirect, is performed by a much less precise algorithm, which also has a branching structure. This algorithm includes empirical-inductive estimation of the stability of subsample averages $M_{k,n}[Y]$. Such multisample estimation is essentially an elaboration of a model of the object of measurement. The result of this elaboration, just as in any other case, may turn out to be negative, i.e., it could show that the concept of the true value of the measured quantity is inadequate in the given specific experimental circumstance. Here this result means that the subsample averages are unstable and, correspondingly, a definite meaning cannot be given sensibly to mathematical expectation.

The uniqueness of mathematical expectation as a measured quantity lies only in the fact that the elaboration of a model of the object of measurement reduces here to realization of the fundamental metrological precept 2°—in the form of estimation of the reproducibility of subsample averages. When other physical quantities are measured, the construction of a model of the object of measurement is usually a more specific experimental-theoretical operation. True, if in measuring the mathematical expectation analysis and selection of the experimental conditions U which are capable, in the opinion of investigators, of providing statistical stability are included in the construction of the model, then the difference vanishes completely (it was pointed out above that a multisample experiment gives a final evaluation of the degree of uniformity of the trials).

We must underscore the metrological ordinariness of measurements of mathematical expectation because of the fact that Fisher's mathematical statistics is abstracted from the need to construct a model of the object of measurement. It assumes that the mathematical expectation always exists and it can always be evaluated from the results of a one-sample experiment.

Sometimes the aim is to measure as accurately and reliably as possible the mathematical expectations of the input or some intermediate quantities of the theory (and not only of the output quantity Y). Everything said above about measurement of the mathematical expectation also pertains in toto to such cases, because here the experiments are essentially of the same character as verifying experiments.

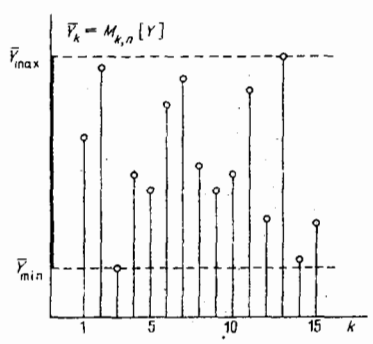

FIG. 4. The multisample confidence interval $[Y_{max}, Y_{min}]$ refers to a set Q of samples (in this case $Q = 15$). (Ref. 29)

5. MULTISAMPLE AND FISHER CONFIDENCE INTERVALS

5.1. Multisample confidence interval

The confidence interval is the key concept of Fisher's mathematical statistics. Utilizing the foregoing considerations, we compare this interval with the interval $[\bar{Y}_{min}, \bar{Y}_{max}]$ in Eq. (4.9), arising in multisample estimation of the reproducibility of averages. The latter interval can be termed a multisample 100% confidence interval for statistical averages. If it is sufficiently small, a numerical value $M[Y]$ of the mathematical expectation is extracted from it. In such a favorable case, this interval can be called the confidence interval for mathematical expectation also. In this sense, measurement of a multisample confidence interval $[\bar{Y}_{min}, \bar{Y}_{max}]$ is a stage in the process of verification of a probability-theoretic prediction $U \to M[Y]$.

A multisample confidence interval is shown in Fig. 4 (thick line along the Y axis). The secondary sample (4.6) shows a grid plot of the dependence $\bar{Y}_k = M_{k,n}[Y]$, $k = 1,...,Q$. If so desired, one can introduce, say, a 90% multisample confidence interval, cutting off 5% of the total number of points at the top and bottom of the grid plot.

In accordance with the antischolasticism thesis of natural science being presented (the validity of a theory is ultimately based on experiment and not on another theory), an ideal account of measurement of a multisample confidence interval could consist of the following three points (Ref. 12, pp. 49; Refs. 13–18):

1) Q subsamples of size n were obtained under controllable experimental conditions U (these conditions must be described in detail);

2) it was found that the subsample averages $M_{k,n}[Y]$, $k = 1,...,Q$, fall within a definite interval $[\bar{Y}_{min}, \bar{Y}_{max}]$; an empirical-inductive prediction is made that they will also fall into this interval for all $k > Q$;

3) the nearest precedents are: ... here it is desirable to compare with previous measurements of a multisample confidence interval under conditions analogous to U. If these measurements were performed in the process of verifying theories, then the practical results achieved with the help of the statistical predictions adopted must be reviewed.

For example, in the case when a physical theory is being checked, it is expedient to confirm the data obtained by different groups of investigators or indicate positive (negative) results obtained in related experiments. In the case of applied theories, concerning technical setups, one may be dealing with information concerning fault-free operation or, conversely, breakdowns.

We point out that no probability-theoretic hypotheses enter into such a report. If the verification of the probability-

theoretic prediction $U \to M[Y]$ is made on the basis, once again, of a probability-theoretic model, then logical cycling obtains: This model must then be verified. After all, as soon as it is acknowledged that the primary probability-theoretic model giving the prediction $U \to M[Y]$ must be verified, there are no grounds for taking, under the conditions U, the secondary probability-theoretic model "at its word."

5.2. Fisher confidence intervals

We now consider the Fisher confidence interval for the mathematical expectation $M[Y]$. In practice, the interval

$$[\bar{Y}_N - t(N,\mathscr{P})S_N,\ \bar{Y}_N + t(N,\mathscr{P})S_N] \tag{5.1}$$

is calculated from the entire available sample $\{Y(s)\}_1^N$; here $\bar{Y}_N \overset{\text{not}}{=} M_N[Y]$ is the sample average, $\mathscr{P}$ is the chosen value of the confidence probability,

$$S_N \overset{\text{not}}{=} [\sum_{s=1}^{N} (Y(s) - \bar{Y}_N)^2 (N-1)^{-1}]^{1/2} \tag{5.2}$$

is the sample rms deviation, and $t(N,\mathscr{P})$ is Student's coefficient.

Users of Fisher's mathematical statistics usually assume that the computed specific confidence interval (5.1) covers the unknown mathematical expectation $M[Y]$ with the probability $\mathscr{P}$. Metrological standards[5] require precisely such an interpretation. For example, according to the state standard GOST 8.207-76 the interval (5.1) must be taken as the interval "in which the total error of measurement falls with an established probability."

Meanwhile, such an interpretation of the interval (5.1) is meaningless, even purely syntactically. After all, the specific interval (5.1), calculated for a single sample, either does or does not cover the unknown value of the mathematical expectation $M[Y]$ as soon as it is judged to exist. Thus here the probability of covering is either unity or zero, and it cannot equal $\mathscr{P}$, if $0 < \mathscr{P} < 1$.

Actually, Fisher confidence intervals of the type (5.1) still make a certain sense. Suppose that $Q \gg 1$ subsamples $\{Y_{(k)}(s)\}_1^n$, $k = 1,...,Q$, of size n have been extracted from a normal universe with the help of independent trials. Suppose further that for each subsample a Fisher confidence interval has been calculated:

$$[\bar{Y}_{k,n} + t(n,\mathscr{P})S_{k,n},\ \bar{Y}_{k,n} + t(n,\mathscr{P})S_{k,n}],\quad k = 1,...Q; \tag{5.3}$$

here $\bar{Y}_{k,n} \overset{\text{not}}{=} M_{k,n}[Y]$ is the subsample average and $S_{k,n}$ is the subsample rms deviation, calculated using a formula similar to (5.2). Under the assumptions made the inequalities

$$\bar{Y}_{k,n} - t(n,\mathscr{P})S_{k,n} \leq M_{k,n}[Y] \leq \bar{Y}_{k,n} + t(n,\mathscr{P})S_{k,n} \tag{5.4}$$

will be valid approximately for $\mathscr{P} \cdot Q$ "good" random intervals (5.3), while for the other, "bad" intervals these inequalities are violated. More precisely,

$$\text{Prob}[\bar{Y}_{k,n} - t(n,\mathscr{P})S_{k,n} \leq M_{k,n}[Y] \leq \bar{Y}_{k,n} + t(n,\mathscr{P})S_{k,n}] = \mathscr{P}. \tag{5.5}$$

The theory contains no indications of which random intervals from the set (5.3) are good.

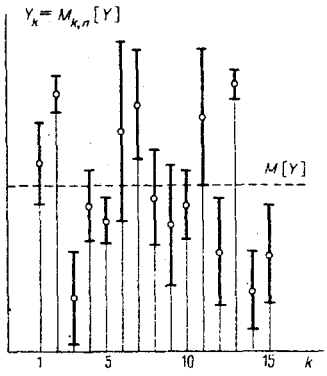

FIG. 5. Fisher confidence intervals for the data presented in Fig. 4. This figure shows that in the present example the multisample average $M[Y]$ falls within the Fisher confidence intervals only for half of all samples. It is this circumstance that indicates the trouble in Fisher's procedure for determining confidence intervals.

This is illustrated in Fig. 5, taken from Ref. 29 (p. 235). As in Fig. 4, the circles indicate the values of the subsample averages. The Fisher confidence intervals are represented by the thick vertical bars. These confidence intervals were found for $\mathscr{P} = 0.5$, $Q = 15$. The horizontal dashed line indicates the value of $M[Y]$.

Figures 4 and 5 illustrate the important difference between a multisample confidence interval and the Fisher confidence intervals: The former interval is measured once for the entire existing set of subsamples, while there are as many Fisher confidence intervals as there are subsamples.

One can see from Fig. 5 that the inequalities (5.4) are valid for approximately half of the Fisher confidence intervals. Here the value of $M[Y]$ turned out to be known, probably because it was incorporated beforehand in the normally distributed pseudorandom number generator, with whose help the subsamples were simulated. In a real experiment the value of $M[Y]$, naturally, is unknown. We recall also that it is unknown beforehand whether or not the concept of mathematical expectation is applicable in a given situation.

The randomness of the position and size of the Fisher confidence intervals, calculated for a set of subsamples from the same universe, is indicated in many sources. However, a clear graphical illustration that would help in gaining a complete understanding of these concepts is rarely presented.

It is instructive that Ref. 9, a handbook on applied statistics, which contains a large number of illustrative computational examples, does not give any examples in the section devoted to Fisher confidence intervals. The authors did not undertake a calculation of an individual Fisher confidence interval in the handbook, probably because such a calculation is meaningless. A set of confidence intervals (5.3) for several subsamples was not calculated in the handbook probably so as not to perplex users of the handbook: It is not clear what one should do with a set of confidence intervals such as the set presented in Fig. 4 (moreover, State standards do not prescribe the calculation of a set of Fisher confidence intervals).

The thought process leading to the incorrect interpretation of the Fisher confidence interval as some fixed interval into which something falls with probability $\mathscr{P}$ can be followed, for example, in Ref. 30. First, one considers the probability

$$\mathscr{P} = \text{Prob}(Y_1 \leq Y \leq Y_2) = \int_{Y_1}^{Y_2} w(Y)dY \qquad (5.6)$$

that a random quantity Y, having a probability density $w(Y)$, falls into the fixed interval $[Y_1, Y_2]$. Into formula (5.6) one can indeed substitute any specific numbers $Y_{1,2}$ and it does not become meaningless [in Ref. 30 the probability $\mathscr{P}$, calculated using the indisputable formula (5.6), is also called the confidence probability]. Next, the relation (5.5) [in Ref. 30, Eq. (36)] is considered to be completely analogous to the formula (5.6) and the specific numerical values of $\overline{Y}_N$ and S_N, found for the available sample $\{Y(s)\}_1^N$, are substituted into it. Ultimately, in Ref. 30 relations of the type (5.5) are replaced, for example, by the equality

$$\text{Prob}(31{,}0 \leq M[Y] \leq 31{,}4) = 0{,}86 \qquad (5.7)$$

[the notation is ours—authors], which is meaningless, since $M[Y] = \text{const}$.

A simpler example of such a mixup is as follows: If a specific realization of a random number Y, say, $Y = 31.0$, is substituted into the inequality $\text{Prob}(Y < \mu) = \mathscr{P}$, $0 < \mathscr{P} < 1$, where μ and $\mathscr{P}$ are constants (possibly, unknown), then the meaningless result $\text{Prob}(31.0 < \mu) = \mathscr{P}$ is obtained. In reality, the number 31.0 is or is not less than the constant μ, so that the probability that the inequality $31.0 < \mu$ is satisfied is equal to unity or zero, but certainly not to the number 0.86.

In considering the Fisher confidence intervals, we encountered a characteristic feature of the Fisher mathematical statistics. On the other hand, the basic concepts of this theory are quite sensible, similarly to confidence intervals, only in application to many subsamples. They are not applicable literally to one sample, though, alas, this is done. On the other hand, the interpretation which Fisher concepts are given in the case of many subsamples is not likely to be suitable for the user. Thus, the entire set (5.3) of Fisher confidence intervals is not likely to be suitable for anyone. It is in this that the above-mentioned inadequate adaptability of Fisher mathematical statistics to a multisample scheme of a verifying experiment manifests itself.

5.3. Extralogical components of Fisher statistics

We now sum up our results. The calculation of Fisher's confidence interval (FCI) for $M[Y]$ is based on the following complex of probability-theoretic hypotheses (we denote this complex by H):

A) the quantity Y is random, so that its mathematical expectation exists;

B) moreover, the probability distribution of the random quantity Y exists and it has a definite form: normal for the continuous random quantity Y, binomial for the binary indicator-quantity $Y = I_A$;

C) the tests are independent in the probability-theoretic sense.

As a result, the verification of the prediction $U \to M[Y]$ separates into three components, which we designate for clarity by the symbols $U \Rightarrow H$, $H \Rightarrow \text{FCI}$, and $\text{FCI} \Rightarrow M[Y]$. The arrow $\Rightarrow$ indicates a formal logical implication operation. The adoption of the concept of a complex of hypotheses H on the basis of analysis of the conditions U is extralogical. The Fisher confidence interval in the component $H \Rightarrow \text{FCI}$ is calculated on the basis of a formal model. The adoption of a final estimate for $M[Y]$ on the basis of the computed FCI (the operation $\text{FCI} \Rightarrow M[Y]$) is once again extralogical, at least because the adoption of the value $\mathscr{P}$ of the confidence probability is extralogical.

In short, we have not, of course, escaped from extralogical components, but rather we have merely wedged between them a formal-logical conclusion from the by no means trivial premise H. The premise itself must be verified, and it is much more difficult to verify than the prediction $U \to M[Y]$ (which, by the way, is the only one of interest to us in the context under consideration). Moreover, the procedure for verifying correctly the premise H will necessarily include the measurement of $M[Y]$. This must be performed according to the branching procedure described above (see Fig. 3) already when verifying the hypothesis A. After such a measurement one should stop, since the prediction $U \Rightarrow M[Y]$ would then, strictly speaking, have turned out to be verified. Meanwhile, the hypotheses B and C... would remain unverified.

With regard to verification of the hypothesis B we note the following. Suppose that we are required to establish reliably a measure of the agreement between the statistical distribution of the sample $\{Y(s)\}_1^n$ and a hypothetical distribution, which we assume to be the normal distribution

$$w(Y) = \frac{1}{\sqrt{2\pi}\sigma[Y]} \exp[-(Y - M[Y])^2/2\sigma^2[Y]]. \qquad (5.5)$$

According to common sense this problem cannot be solved without first estimating for the sample $\{Y(s)\}_1^N$ as accurately as possible the parameters $M[Y]$ and $\sigma[Y]$ (the off-set and the scale, respectively) of the distribution (5.5). This means that the hypothesis A must still be verified first (but then, once again, the job is already done, and the hypotheses B and C will be superfluous).

Fisher's mathematical statistics does not take into consideration this natural hierarchy of estimation problems. We can say that it proposes checking somehow the similarity between the above-mentioned sample and hypothetical distributions, and then it considers it acceptable to base on the results of such a rough check the further quite complicated deductions of the theory of Fisher confidence intervals. We offer, in this connection, the words of I. Grekova (the literary pseudonym of E. S. Venttsel', a well-known specialist in applications of the theory of probability in aviation): "A quite subtle apparatus, based on the assumption that the distribution of the observed random quantity (normal distribution) is known, has been developed for calculating FCI. And again there arises the question: And on what basis, strictly speaking, is this known? How accurately is it known? Finally, what is the practical value of the "product" itself—the confidence interval? Few experiments means little information, and we are in a bad predicament. Whether or not the confidence interval in such a case is a little larger or smaller is not so important, especially since the confidence probability is assigned arbitrarily" (Ref. 31, p. 111) (the last phrases in this statement allude to the fact that the theory of FCI is oriented primarily toward a "miscalculation" involving small samples).

To I. Grekova's statement we can possibly only add a reminder that an individual Fisher confidence interval (5.1)

generally does not have the meaning which users of mathematical statistics seek in it.

The hypothesis C is the most difficult one to check, at least for the reason that in determining the probability-theoretic independence both multivariant and univariant probability distributions are present. Moreover, it is much more difficult to check the independence of trials than it is to give a probability-theoretic proof of the independence of random quantities (see, for example, Ref. 12).

A more detailed critical analysis of Fisher confidence intervals and the principles of mathematical statistics was presented in Refs. 12–18, 23, and 24. The programs for measuring $M[Y]$, presented in Sec. 5.1 (points 1–3), offer a constructive alternative to Fisher mathematical statistics.[12–18,23,24]

6. ARE THERE ENOUGH DATA FOR A RELIABLE PREDICTION?

6.1. Incompleteness of the system of hypotheses

We shall list and, where possible, comment on other extralogical relations brought into the calculation of probabilities. We begin with the general question of the role of hypotheses in probability calculations.

Any system of hypotheses is incomplete. This incompleteness principle finds many confirmations in life: No matter how far-sighted and scrupulous we are in constructing scenarios of the behavior of a complicated system, there are still many (even infinitely many!) factors which can affect the resulting behavior.

An example of incompleteness of hypotheses are virtually all the defects appearing in nuclear power plants, space systems, and other complicated technological devices. In estimates of the reliability of a nuclear power plant, one would think that all imaginable reasons for malfunctions and accidents are accounted for. Estimates give such low probabilities for malfunctions that one can only wonder at how malfunctions arise at all.

But the entire point is that the calculation of probabilities is always based on an *incomplete* system of hypotheses. The fire at one of the American nuclear power plants is an instructive example which P. L. Kapitsa gives: The fire was blamed on an electric lamp that burned out in a room where a leak occurred in a water faucet and the metal worker could find nothing better to do than to light a candle in a dark room and thereby created the focus of the fire.

In the construction and operation of the Chernobyl nuclear power plant it hardly occurred to anyone to include among significant factors the incompetence of the personnel and their inability to understand the amorality of unsanctioned experiments. Incidentally, this is not the only example when the moral responsibility is just as important (if not more important) a safety factor as the technical characteristics of complicated systems. How short does the path turn out to be from moral principles to the operational reliability of complicated systems! And how spontaneously parallels arise with the activity of previous criminal structures in the Union which could not waive their principles... .

6.2. Subjective estimates of probabilities

Subjective (expert) estimates of probabilities are most often resorted to when there are too many indeterminate (in Tutubalin's sense) factors, which cannot be inserted either into a determinate or a statistical model of the prediction.

Subjective estimates are based on the previous experience of an expert, which, in practice, often cannot be formalized. As a rule, it is pointless to subject such estimates to verification at the level adopted in natural science. The practical results of economic, technical, etc. activity are most likely subjected to verification. The criteria for such verification are usually unclear, just as, by the way, the forecasting estimates of experts, and it is difficult to include them in the natural-science paradigm. This is why no success has been achieved (and is it necessary?) in constructing a bridge between the natural sciences, on the one hand, and astrologers and psychics on the other.

6.3. Formal construction of a statistical ensemble

The construction of the statistical ensemble (universe) from a finite experimental sample is one of the most complicated questions in the practical theory of probability.

Suppose that we have measured the statistical characteristics of a process on the interval $[0,T]$. The entire complexity of the construction problem reduces to the formal construction of the statistical characteristics outside of the interval $[0,T]$. In most cases the simplest method is employed—the principle "tomorrow will be the same as today" is used, making the assumption that the statistical characteristics for $t > T$ will be the same as for $0 < t < T$.

Of course, no one can guarantee that this will be the case for an infinitely long time, and for this reason where possible and expedient the statistical information must be constantly renewed and the appearance of significant changes must be monitored.

The use of experimentally measured time averages as the characteristics of a statistical ensemble has been termed the ergodicity hypothesis. In essence, the property of ergodicity reflects nothing more than our belief in the validity of using time averages as parameters of a hypothetical statistical ensemble. Any change in the time averages can serve as a signal for reexamining the characteristics of the statistical ensemble.

Of course, this is demanding. But, if instead of a painstaking analysis of the properties of the real system some words are uttered about the ergodicity of the system, then the investigator seemingly obtains an indulgence for the case when possible deviations occur from the adopted hypothesis and is thereby absolved from the need to adjust the imagined statistical ensemble to the changing conditions.

We now examine some dangers facing the investigator "gambling" on a definite statistical analysis.

6.4. Nonstationariness

Any efforts to expose the nonstationariness of the statistical characteristics of a process must be restricted at the outset to the case of slow nonstationariness. The point is that statistical characteristics are nonlocal, they are formed over longer or shorter time intervals.

Suppose a multisample average $M[Y]$ is determined with uncertainty $\Delta\overline{Y}$ over a time T. Then experimentally only nonstationary changes not less than $\Delta\overline{Y}$ are revealed over the time T. Thus the minimum degree of nonstationariness, "decomposed" on the interval $[0,T]$, is

$$\min\left|\frac{dM[Y]}{dt}\right| \sim \frac{\Delta\bar{Y}}{T}. \tag{6.1}$$

At the same time the observation interval T must be short compared with the time of nonstationary evolutionary change

$$t_{evol} \sim |M[Y]|/|dM[Y]/dt|. \tag{6.2}$$

With the help of the condition $T < t_{evol}$ we obtain from the relation (6.1) the inequality

$$\frac{\min\left|\frac{dM[Y]}{dt}\right|}{\left|\frac{dM[Y]}{dt}\right|} \gtrsim \frac{\Delta\bar{Y}}{|M[Y]|}, \tag{6.3}$$

which means that the error with which the derivative $dM[Y]/dt$ is determined is always greater than the error of measurement of $M[Y]$.

Thus the interval of multisample measurement cannot be too short (in this case, the error $\Delta\bar{Y}$ increases) or too long, for otherwise the nonstationariness effect being sought could be missed. The minimum time for measuring nonstationariness [it is estimated from the relation (6.1) with a prescribed value of $\min|dM/dt|$] restricts the rate of nonstationariness.

On the whole, this question has not yet been adequately elucidated in the literature. We note only that by adopting the hypothesis of ergodicity, we can assign (adjoin) a statistical ensemble not only to a stationary, but also to a nonstationary process.

6.5. Instabilities

Instabilities are especially dangerous for predictions, since they begin to develop out of sight, buried in the noise accompanying any measurement. Initially, after the instability is generated, the rapidly growing exponential is still hidden in the noise. As a result, the growth of an instability becomes observable only sometime after the instability exceeds the noise level. After this, the unstable process continues to grow rapidly and reaches macroscopic values over a finite time, sometimes literally within several measurements. The statistical characteristics of the process under investigation correspondingly also change.

The discovery of instabilities of one or another nature beforehand is a problem of enormous importance in many fields of science and technology, for example, in the problem of controlled thermonuclear fusion. Sometimes the instability is not exponential, but rather explosive, and then the growth time of the instability can be even shorter. Finally, we mention also processes such as earthquakes, which arise as a result of the accumulation of static stresses and are manifested in the form of short-duration releases of stored energy.

6.6. Rare phenomena

In many problems there is no hope not only for reproducibility, so as to be able to verify the statistical characteristics of the process or phenomenon under study, but even for a single repetition of an observation. We are talking about some natural phenomena occurring in space (supernova), in the ocean (unusual flows), and in the atmosphere (rare optical and weather phenomena), as well as rare phenomena occurring in laboratory experiments (detection of rare transformations in high-energy physics, detection of high-energy cosmic rays).

In such situations statistical predictions are customarily constructed according to the ethical rules of physical experimentation, i.e., maximum self-criticism is exercised and all alternative hypotheses are considered. It is due to these unstated rules that even the data and hypotheses that engender legitimate doubts are usually analyzed in a quiet atmosphere with the desire to achieve maximum objectivity. This is how, or approximately how, the hypothesis of the existence of a neutrino rest mass, which was later not confirmed (but not rejected either!), was discussed.

Of course, sad exceptions also occur, examples being the case of publications on cold nuclear fusion and on the biological effect of pure water.

6.7. "Side dishes"

While within the framework of the modern scientific paradigm there is sufficient strength to maintain "ethical health" in verifying even rare and unique phenomena, in the quasiscientific environment, centered on UFOs, telekinesis, the Bermuda triangle, and psychics, assertions are very often encountered with regard to which the term "pseudoscience" will not seem exaggerated.

We are talking not so much about the boldness of the hypotheses under discussion (bold hypotheses are by no means rejected by modern science), but rather about the *level of discussion of them*. It is precisely the unrestricted treatment of hypotheses and complete ignoring of the ethical rules of experimentation that relegates pseudoproblems to the distant periphery of natural science, to the level of medieval thought.[1)] Of course, the problems of metrological support, verification of hypotheses, and probabilistic interpretation do not arise here at all—for the conscientious natural scientist there is simply nothing to verify here.

6.8. Classical probabilities

Probabilities corresponding to the classical definition (ratio of the number of favorable outcomes to the total number of possible outcomes), from the modern viewpoint, standout simply as the simplest hypotheses concerning the frequency of appearance in idealized systems—a coin, a die, etc.

Depending on the conditions under which an experiment is performed, real frequencies of appearance can differ from the classical frequencies. Thus the actual results depend on the positions of the center of gravity of the die, the area of the faces of the die, the smoothness of the angles, the quality of surface of the table, and other physical characteristics of the object and the experiment. In addition, as J. Keller recently analyzed for the example of flipping a coin, the result depends on the initial linear and angular velocities of the flipping.[33] It was found that heads and tails are obtained with equal probability only asymptotically for high initial velocities.

Thus even classical probabilities sometimes require experimental verification, especially if a game is suspected to be dishonest.

6.9. Fiction of the law of large numbers

In the physical literature there still reigns the belief that as the number of trials (the sample size) N increases because

of the central limit theorem the relative frequency of appearance approaches its limit, which is the empirical probability. Meanwhile, experiments very often indicate that for large N the dispersion of the data does not decrease. On the contrary, at some value N_0 it increases. A more detailed and serious exposition of this problem is given by P. E. El'yasberg.[26]

Without being too precise, the reason the dispersion increases is that for $N > N_0$ the systematic error arising due to the fact that the base hypothesis does not include some significant factor which, at first, for a small sample size, would give an insignificant contribution and then, as data are accumulated, would become increasingly more noticeable.

Thus if the model of a phenomenon does not include any significant systematic factor, then the dispersion will not necessarily decrease with increasing sample size.

Another possible reason for the fact that the dispersion does not decrease (or does not decrease sufficiently rapidly) could be the character of the fluctuations. The conditions, presupposed by the central limit theorem, though they are not too burdensome, still are not always automatically satisfied. For this reason, verification of the conditions of applicability of the central limit theorem is, in many cases, not only desirable, but simply a necessity.

We can thus state that a decrease of fluctuations with increasing N, viewed as a physical fact, is not a trivial consequence of the central limit theorem, but rather it occurs only under certain definite conditions which must be specially checked.

7. ALGORITHMIC COMPLEXITY AND PARTIAL DETERMINATENESS

7.1. Physical experiment and the concept of algorithmic complexity

A system of conventions arises not only in the measurement of probabilities, but even at a still earlier stage—at the stage of definition of the concept of randomness. The set-theoretic approach refers to those quantities as being random, that are equipped with a probability measure. The applied theory of probability distinguishes the class of random quantities according to the stability of the statistical characteristics.

The algorithmic theory of probability[34,35] identifies randomness with algorithmic complexity. Finally, randomness is interpreted in the theory of partially determinate processes as unpredictability.[36] As we can see, even in the question of what should be termed random, there are at least several conventions (a more complete list is given in Ref. 36).

We have already discussed above the relation between the empirical and set-theoretic averages. We now briefly discuss the relation to the algorithmic theory of probability and the theory of partial determinateness.

Contrary to expectation, the algorithmic theory of probability strictly speaking is not concerned with the calculation of probability. The problem addressed by the algorithmic theory is to establish a criterion for randomness, interpreted as the algorithmic complexity of a sequence of numbers (the results of measurements are a sequence of numbers). In the context of this paper, the algorithmic theory of probability is interesting in that it explicitly relates randomness to algorithms, i.e., ultimately hypotheses and formal constructs.

Complexity is defined as the shortest length $l_{\min}$ (number of operations) of an algorithm that converts one sequence of numbers $\{x\}$ into another sequence of numbers $\{y\}$. We note that here there arises a new condition, associated with the existence of a *set* of hypothetical algorithms that convert $\{x\}$ into $\{y\}$ and with the need to search for the shortest algorithm.

If the length $l(N)$ of an algorithm remains finite in the limit $N \to \infty$, more precisely, if

$$\frac{l(N)}{N}\underset{N\to\infty}{\to} 0, \tag{7.1}$$

then the sequence is declared to be algorithmically simple and, consequently, not random. If, however, $l(N)/N \to \text{const}$, then the sequence is *algorithmically complex* and thereby *random*.

In spite of the attractiveness of the concept of randomness as algorithmic complexity, proposed by A. N. Kolmogorov[34] and developed by his followers, on the whole it is unlikely to be of interest for the problem of physical measurements. First of all, this concept invloves the limit $N \to \infty$ and requires the performance of tests, which become longer and longer, in order to reveal first simple and then more and more complicated algorithms. Such a sequence of tests of increasing length should ultimately approach Martin-Löf's "universal test."[35] It is obvious that such an operation of testing for all imaginable and unimaginable algorithms is unrealizable in practice, while the concept of algorithmic complexity itself, in principle, requires passage to the limit $N \to \infty$.

Second, the concept of randomness as algorithmic complexity contradicts the view of natural phenomena as it is developed in physics. It is difficult for the naturalist to acknowledge a sequence (process) as being random if its algorithm is known, even though it is complex.

Third, there exists a hardly surmountable difficulty associated with the presence of noise in measurements. In physical measurements, as a rule, noise is filtered (discriminated), otherwise the noise and not the process under study will be the object of measurement. In the algorithmic approach, if it is applied to measurements, noise is not distinguished from the measured process; in any case we do not know of any such attempts. Ultimately, even with a relatively simple algorithm of the process under study, the mixture "signal + noise" acquires the complexity of the noise.

Thus the logic of a physical experiment is, in a significant way, hardly compatible with the concept of algorithmic complexity.

7.2. Physical experiment and the concept of partial determinateness

In the part concerned with determining the dynamical (determinate) laws in the process under study, the logic of an experiment is closer to the concept of partial determinateness.[36] The latter concept judges the observed process $y(t)$ to be determinate or random according to the degree to which the process is similar to a model predicting process $z(t)$.

The degree of determinateness[36]

$$D(\tau) = \frac{\{y(t)\cdot z(t)\}}{(\{y(t)\cdot y(t)\}\{z(t)\cdot z(t)\})^{1/2}}, \tag{7.2}$$

where $\tau = t - t^0$ is the time which has elapsed from the start

of observations and the braces indicate the operation of comparing (projection), can serve as a quantitative characteristic of coincidence. In comparing the observations $y(t)$ with a model process $z(t)$ it is presupposed that the initial conditions for $z(t)$ are the same as for the recorded process $y(t)$, namely,

$$z(t^0) = y(t^0). \tag{7.3}$$

Because of this, at $\tau = 0$ the degree of determinateness (7.2) is equal to unity, however the comparison operation $\{\cdot\}$ is defined.

The equality $D = 1$ corresponds to *complete determinateness* (complete predictability) of the observed process $y(t)$ relative to the model process $z(t)$. In the opposite case, vanishing of D is interpreted as *complete randomness* (unpredictability) of $y(t)$ with respect to $z(t)$. The values of D such that $0 < |D| < 1$ describe *partial determinateness*. The interval during which the degree of determinateness D exceeds a certain level, say, $D \geqslant 1/2$, characterizes the time of the determinate behavior or, which is the same thing in the present context, the time of predicability τ_{pred}.

In the original publications[36,37] the operation of statistical (empirical) averaging of the product $y(t)z(t)$ was chosen as the comparison operation:

$$\{y(t)\cdot z(t)\} = \langle y(t)z(t)\rangle, \tag{7.4}$$

so that the degree of determinateness D is the correlation coefficient between $y(t)$ and $z(t)$. Statistical averaging can be combined with the integration over time,

$$\{y(t)\cdot z(t)\} = \int_{t^0}^{t^0+\tau} \langle y(t)z(t)\rangle dt, \tag{7.5}$$

and then the measure (7.2) characterizes not the local, as does the relation (7.4), but rather the *integral* coincidence between the observation and the prediction over the entire segment $[t^0, t^0 + \tau]$.

The concept of partial determinateness formalizes the actually existing relations between the experimenter and the experimental material. The experimenter puts forth hypotheses $z(t)$ and checks the experimental data $y(t)$ against them. In the real world no model $z(t)$ can claim prediction over an infinitely long time, and for this reason for *any physical processes*

$$\tau_{\text{pred}} < \infty.$$

This gives us an opportunity to look at the algorithmic approach from a somewhat different viewpoint. First of all, from the physical standpoint it is pointless to subject the testing process to infinitely long tests—it is sufficient to confine the tests to finite times $\tau < \tau_{\text{pred}}$.

Second, there is no need to pursue universality (according to Martin–Löf, these are all imaginable tests and all tests that future generations of people can suggest). It is much more practical to consider existing tests ($\equiv$ hypotheses, models), i.e., to estimate the degree to which the observations $y(t)$ coincide not with all imaginable hypotheses, but only those hypotheses which are actually available to the experimenter.

Finally, the concept of partial determinateness radically solves the problem of noise, which is always present in the observation $y(t)$: The comparison operation (7.5) includes *filtering*. The specific weight of the components of the observation $y(t)$ which originate from the noise decreases with a prolonged accumulation with weight $z(t)$, just as, by the way, do the components of $y(t)$ which do not agree with the adopted model $z(t)$. For this reason, repeated accumulation of data makes it possible to separate the signal [i.e., "comprehended" part of $y(t)$] from the noise. In the algorithmic approach, however, as we have seen above, noises are not subjected to sensible filtering. With these significant additions, the concept of partial determinateness can be viewed as an elaboration of the algorithmic approach to real physical objects of investigation.

7.3. Empirical probability as the degree of determinateness

The concept of partial determinateness turns out to be very flexible and universal. As evidence of this flexibility, we point out that in defining the operation $\{y(t)\cdot z(t)\}$ as the *number of coincidences* between the values of $y(t)$ and $z(t)$ the degree of determinateness (7.2) becomes the *empirical probability*.

In order to verify this, we discretize the readings both in time (the sth reading is taken at the time $s\Delta t$ after t^0) and in magnitude: The values of y and z are taken with a sampling interval ε. We define the comparison operation $\{y\cdot z\}$ as the number of ε coincidences between y and z within a strip of width ε. It can be expressed as the number of events A, which consist of the fact that the modulus of the difference $y - z$ does not exceed $\varepsilon/2$:

$$\{y\cdot z\} = n_A(\tau), \text{ where } A\colon |y - z| < \varepsilon/2. \tag{7.6}$$

If the indicator function $I_A(s)$, defined by the relation (4.3), is introduced, the number of ε-coincidences $n_A(\tau)$ in the interval $[t^0, t^0 + n\Delta t]$ is given by the sum

$$\{y(t)\cdot z(t)\} = n_A(\tau) = \sum_{s=1}^{n} I_A(s), \quad \tau = n\Delta t. \tag{7.7}$$

Since for $y = z$ we always have $I_A(s) = 1$ and $\{y\cdot y\} = \{z\cdot z\} = n$, the degree of determinateness (7.2) becomes the ratio of the number of coincidences $n_A(\tau)$ to the total number of readings performed over the time $\tau = n\Delta t$:

$$D(\tau) = \frac{n_A(\tau)}{n} = \frac{\sum_{s=1}^{n} I_A(s)}{n} = \omega_n(\varepsilon). \tag{7.8}$$

The relation (7.8), as one can easily see, is the relative frequency $\omega_n(\varepsilon)$ of coincidences between y and z within the strip ε, i.e., it is the empirical probability p.

In spite of the existence of a direct relation between $D(\tau)$ and the sample probability p, the quantity D, as a measure of the degree of coincidence between y and z, still has some additional flexibility, consisting of the fact that in the comparison process the values $z_s = z(t_s) = z(t^0 + s\Delta t)$ can change together with the observed quantity $y_s = y(t_s) = y(t^0 + s\Delta t)$. If the model (predicted) value is constant, $z = z_*$ [in this case the requirement (7.3) on the initial condition $z(t^0)$ must be removed], then the quantity (7.8) is simply the relative time (relative frequency) the observed quantity $y(t_s)$ occupies an ε-neighborhood of the fixed value z_*:

$$D(\tau) = \omega_* = \omega(|y - z_*| < \varepsilon/2). \tag{7.9}$$

Now suppose that we can predict the value y_s, i.e., there exists a satisfactory law (rule, algorithm, guess, secret information, etc.) for constructing the prediction z_s. Then an ε-coincidence between all terms of the sequences

$$y_1, y_2, \ldots, y_n \text{ and } z_1, z_2, \ldots, z_n \tag{7.10}$$

will mean that

$$D(\tau) = 1. \tag{7.11}$$

Comparing the values (7.9) and (7.11) we can see that there seemingly exist hidden conditionalities in the simplest operations, such as counting the number of events. We are talking about comparing the frequencies of a definite fixed event, say, obtaining a six in the case of a die, with the frequency of coincidence between the observed quantity y_s and the (variable) prediction z_s. Continuing the example of the die, in the case of an "honest" die, for a sufficiently large number of rolls n the frequency of obtaining a six ($z_* = 6$) will approach 1/6. This means that the numbers in the observed sequence $y_1, \ldots, y_n$ will coincide with $z_* = 6$, on the average, one out of six times.

The situation is different in the case of a die that is controlled, for example, with the help of magnet, when the manipulator organizes the values y_s known to him. If these values are taken from a table of random numbers, then coincidence with the fixed value $z_* = 6$ will be observed, as before, in approximately 1/6 of the throws. An objective (or, more accurately, honestly naive) observer, making judgments assuming an "honest" die, will say that the results of the rolls correspond to his intuitive expectations. At the same time, the manipulator, i.e., the informed observer, for whom the result y_s of rolling the die is known beforehand, will predict the result based on the "dishonest" model $y_s = z_s$, where z_s is the number which he himself has set. In this case

$$D(\tau) = 1.$$

In the intermediate case, when the manipulator does not completely control the result, the quantity D can fall between the values $\omega_* = 1/6$ and $D = 1$.

Manipulation can also consist of using a "dishonest" die, for which $\omega_* > 1/6$. Then the strategy of a "dishonest" player will reduce to predicting a six more frequently. It is interesting that in the case of false information, obtained by the manipulator from the controlling apparatus, the predictions can give values $D < 1/6$. For example, if the apparatus is "harmful," i.e., it always gives what is not required of it, then in general $D = 0$.

We intentionally used the example of games of chance in order to illustrate the main idea. The point is that the work performed by the naturalist is in many ways similar to that of a decipherer, who tries to extract from the set of received messages laws which operate in nature. The experimenter interprets the obtained information on the basis of hypotheses. The example we have just considered shows that the object demonstrating "random" behavior (the probability of obtaining a particular face of a die is equal to 1/6), can still obey some algorithm, a determinate law of behavior.

Here we once again encounter multiple meaning of terms. In the present case, *determinateness*, interpreted as the compliance of a process to a more or less complicated law, does not contradict *randomness*, interpreted as equal probability of obtaining any face of a die.

On a general level, this indicates a profound difference between *internal* probability characteristics of a process $y(t)$ (mathematical expectation $M[y]$, higher order moments $M[y_i y_j y_k \ldots]$) and *external* characteristics, revealed by comparing $y(t)$ with the model processes $z(t)$.

As strange as it may seem, physicists are interested not only in internal (in the sense indicated above), but very often also in external characteristics, i.e., they are interested in the *degree of correspondence* between the observation and a model process (theory). The measure of comparison discussed here (7.2) [with the correlation (7.4), integral (7.5), or probabilistic, more accurately, "coincidence" (7.7) operations of comparison] meets exactly the requirements of an experiment.

In the interpretation of an experiment the comparison operation always presumes the existence of a hypothesis, model, and even guesses (the extralogical, according to Feinberg, principle in natural science). This corresponds completely to the central theme of our paper—the most important role of hypotheses and conjectures in analyzing the data of an experiment. We recall also that it is precisely the "Art of Guessing" (Ars conjectandi) that Jacob Bernoulli called his book, the first book in the world on the theory of probability.[38]

8. CONCLUSIONS

We now list the basic results of our analysis:

1) The relative frequency of appearance ω_N, interpreted as the statistical probability, as well as the statistical mathematical expectation $M_N[Y]$ or $M_{k,n}[Y]$ are *normal physical quantities* (i.e., they can be measured and their error can be estimated), when we are dealing with random (in Tutubalin's sense) quantities, which, in contrast to indeterminate quantities, have stable statistical characteristics.

2) The "abnormality" of empirical probability and mathematical expectation consists of the fact that more than other physical quantities they are burdened with conditionalities and hypotheses, which require special checking (verification).

3) Fisher's mathematical statistics contains hypotheses, which cannot be checked experimentally (or do not stand up to such a check) and are even not subject to syntactic analysis, and this indicates that these statistics are unsound.[2)]

4) The traditional and intuitive acceptable alternative to Fisher's statistics is multisample processing of data (Refs. 12–14; Secs. 4 and 5), which is based on a reasonable minimum number of assumptions.

5) Formal constructs are involved in the practical calculation of probabilities in many cases including:

—listing (and deliberately omitting!) factors affecting the reliability of complicated systems;

—subjective estimation of probabilities;

—reconstruction of a statistical ensemble from a limited experimental sample;

—assumption of ergodicity;

—prediction under conditions of nonstationariness and instability;

—interpretation of rare phenomena;

—use of classical probabilities as models of physical phenomena;

—invoking the law of large numbers for analysis of physical phenomena.

[1)] V. L. Ginzburg's article on this subject in the paper Izvestiya in 1991 is very pertinent.

[2)] It will be interesting to see how much time passes before the Russian authorities correct the State standard recommending Fisher's procedure. Will they manage by the end of this millenium?

[1] E. Wigner, Commun. Pure Appl. Math. **13**, 1 (1960).
[2] A. N. Krylov, *Memoirs* [in Russian], Sudostroenie, Leningrad, 1979.
[3] E. L. Feinberg, Vopr. Filosofii, No. 8, 33 (1986).
[4] E. M. Dushin [Ed.], *Fundamentals of Metrology and Electrical Measurements* [in Russian], Energoatomizdat, Leningrad, 1987.
[5] *Fundamental Standards in the Field of Metrological Support* [in Russian], Goskomstandart, Moscow, 1983.
[6] M. Kendall and A. Stuart, *Distribution Theory, The Advanced Theory of Statistics*, Vol. 1, Hafner Publishing Co., N.Y., 1963 [Russ. transl., Nauka, M., 1966].
[7] M. Kendall and A. Stuart, *Inference and Relationship, The Advanced Theory of Statistics*, Vol. 2, Hafner Publishing Co., N.Y., 1961 [Russ. transl. Nauka, M., 1973].
[8] M. Kendall and A. Stuart, *Design and Analysis, and Time Series, The Advanced Theory of Statistics*, Vol. 3, Hafner Publishing Co., N.Y., 1963 [Russ. transl. Nauka, M., 1976].
[9] S. A. Aivazyan, I. S. Enyukov, and L. D. Meshalkin, *Applied Statistics: Fundamentals of Modeling and Primary Data Processing* [in Russian], Finansy i statistika, M., 1983.
[10] S. A. Aivazyan, I. S. Enyukov, and L. D. Meshalkin, *Applied Statistics: Investigation of Relationships* [in Russian], Finansy i statistika, M., 1985.
[11] S. A. Aivazyan, I. S. Enyukov, and L. D. Meshalkin, *Applied Statistics: Classification and Reduction of Dimension* [in Russian], Finansy i statisika, M., 1989.
[12] Yu. I. Alimov, *Alternative to the Method of Mathematical Statistics* [in Russian], Znanie, M., 1980.
[13] Yu. I. Alimov, in *Semiotics and Information Science* [in Russian], VINITI, M., 1985, No. 24, p. 58.
[14] Yu. I. Alimov, in *Statistical Analysis of Experimental Data* [in Russian], Novosibirsk Electronics Institute, Novosibirsk, 1986, p. 15.
[15] Yu. I. Alimov, *Measurement of the Moments of a System of Random Variables* [in Russian], Ural Polytechnical Institute Press, Sverdlovsk, 1984.
[16] Yu. I. Alimov, *Measurement of Spectra and Statistical Probabilities* [in Russian], Ural Polytechnical Institute Press, Sverdlovsk, 1986.
[17] Yu. I. Alimov, *Prediction of Probability Distributions* [in Russian], Ural Polytechnical Institute Press, Sverdlovsk, 1986.
[18] Yu. I. Alimov and A. B. Shaevich, Zh. Anal. Khim. **44**, 1983 (1988) [J. Anal. Chem. (USSR) (1988)].
[19] P. Whittle, *Probability*, Penguin, Harmondsworth, 1970. [Russ. transl. Nauka, M., 1982.]
[20] R. von Mises, *Mathematical Theory of Probability and Statistics*, Academic Press, N.Y., 1964 [Russ. transl., Gosizdat, M., 1930].
[21] V. N. Tutubalin, *Theory of Probability* [in Russian], M., 1972.
[22] A. N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea Publishing Co., N.Y., 1956 [Russ. original, Nauka, M., 1974].
[23] Yu. I. Alimov, Avtomatika, No. 1, 71 (1978).
[24] Yu. I. Alimov, Avtomatika, No. 4, 83 (1979).
[25] A. M. Yaglom, *Correlation Theory of Stationary Random Functions* [in Russian], Gidrometeoizdat, Leningrad, 1981.
[26] P. E. El'yasberg, *Measurement Information: How Much is Necessary? How Should It Be Processed?* [in Russian], Nauka, M., 1983.
[27] V. N. Tutubalin, *Theory of Probability in Natural Science* [in Russian], Znanie, M., 1972.
[28] N. Wiener, *God and Golem, Inc.*, MIT Press, Massachusetts Institute of Technology, Cambridge, Mass., 1964, pp. 89–91 [Russ. transl. Progress, M., 1966].
[29] N. V. Smirnov and I. V. Dunin-Barkovskiĭ, *Course in the Theory of Probability and Mathematical Statistics* [in Russian], Nauka, M., 1965.
[30] A. N. Zaidel', *Errors in Measurements of Physical Quantities* [in Russian], Nauka, Leningrad, 1895.
[31] I. Grekova, Vopr. Filosofii, No. 6, 104 (1976).
[32] P. L. Kapitsa, *Theory, Practice, Experiment*, D. Reidel, Boston, 1980 [Russ. original, Nauka, M., 1984].
[33] J. B. Keller, Am. Math. Monthly **93**, 191 (1986).
[34] A. N. Kolmogorov, Sankhya Indian J. Statist. Ser. A **25**, 369 (1963); Probl. Peredachi Inf. **1**, 3 (1965); both papers are also contained in A. N. Kolmogorov, *Information Theory and Theory of Algorithms* [in Russian], Nauka, M., 1987, pp. 204 and 215.
[35] P. Martin-Lof, Inf. Control **9**, 602 (1966).
[36] Yu. A. Kravtsov, Usp. Fiz. Nauk **158**, 92 (1989) [Sov. Phys. Usp. **32**, 434 (1989)].
[37] Yu. A. Kravtsov, in *Nonlinear Waves. Vol. 2: Dynamics and Evolution* [in Russian], Nauka, M., 1981, p. 81.
[38] J. Bernoulli, *On the Law of Large Numbers* [Russ. transl., Nauka, M., 1986, Part 4.]

Translated by M. E. Alferieff