Studies in the History of Statistics and Probability.
Collected Translations

vol. 2

V. N. Tutubalin, Yu. I. Alimov

On Applied Mathematical Statistics

Compiled and translated by Oscar Sheynin

Internet: www.sheynin.de

ISBN 978-3-942944-04-5

Berlin, 2011

Contents

Introduction by Compiler

I am presenting translations of some contributions special in that they were devoted to the practical aspect of applied statistics. In any case, an acquaintance with them compels the reader to think about unexpected circumstances. I never met Yuri Ivanovich Alimov, but some decades ago I had attended a short course of lectures at Moscow University delivered by Valery Nikolaevich Tutubalin. I regret that he had no desire to have a look at his previous work. He allowed me to include here (see below in translation) his letter to me explaining his reluctance.

Tutubalin himself [v, beginning of] indicated what prompted him to compile his booklets [i – iii] and, as he reasonably supposed, also served as a catalyst for Alimov [iv]: *the amount of falsehoods arrived at by applying the theory of probability is too great to be tolerated*. He cited Grekova (1976) who had quoted scientific lore which stated that *pure mathematics achieves the probable by proper methods and applied mathematics achieves the necessary by possible means*. The problem therefore reduces to verifying those possible means, to ascertaining the conditions for those means to remain possible.

Tutubalin intended his booklets for a rather broad circle of readers even though he was discussing most serious subjects [ii]. But then, in the first place in [iii], his text included hardly comprehensible statements and an unusual pronouncement on Bernoulli's law of large numbers which should be read together with Alimov's works.

Two of Tutubalin's statements in the same booklet (see my Notes 17 and 18) were no doubt watered down to pass censorship; nowadays, they should have been drastically altered.

Two points ought to be indicated. First, concerning the application of probability to administration of justice see my Note 4 to booklet [i]. Second, Tutubalin [i] overestimated Laplace's influence with respect both to theory and general thinking. I think that Fourier (1829, pp. 375 – 376) correctly described Laplace as a theoretician:

We cannot affirm that it was his destiny to create a science entirely new [...]; *to give to mathematical doctrines principles original and of immense extent* [...]; *or, like Newton,* [...] *to extend to all the universe the terrestrial dynamics of Galileo; but Laplace was born to perfect everything, to exhaust everything and to drive back every limit in order to solve what might have appeared incapable of solution*.

Neither Boltzmann (who cited many scholars and philosophers), nor Poincaré (who regrettably knew only Bertrand) referred to Laplace even once, and Maxwell only mentioned him twice in a very general way.

As to general thinking, Quetelet regrettably overshadowed Laplace's *Essai* by his spectacular but poorly justified announcements and proposals later rejected by German statisticians along with the theory of probability.

Alimov's booklet [iv] is written in bad general style. Witness his original first sentence (altered in translation): ... *mathematicians and*

those who applies it ... The booklet is intended for a much better qualified readership. He indicates the weak points of the attempts to apply probability theory, but his positive recommendations are not sufficiently isolated from the context and the exposition is not at all conducive for easy reading. I only translated parts of his booklet and described much in my own words.

Alimov's criticism of the usual practical aspect of applied mathematical statistics is much more radical than Tutubalin's, suffice it to mention the title of his contribution [iv], and he also over-enthusiastically rejected many chapters of that discipline.

A special comment is warranted by the authors' separation of two understanding of randomness, its narrow mathematical meaning and its more general scientific understanding. This latter is still important; its beginning can be traced to Poincaré (1896/1912, p. 4) who indicated that a very small cause can have a considerable effect which was his main explanation of randomness. His idea (effectively pronounced earlier by several scholars including Maxwell and even by Aristotle) was greatly generalized in the studies of chaotic phenomena which began several decades ago. I provide an example illustrating a mistake made by imagining mathematical randomness instead of randomness in the general sense (or even simply indefiniteness).

William Herschel (1817/1912, p. 579) formulated a statement about the size of the stars. Not knowing anything about it or about the existence of different spectral classes, he *presumed* that a star randomly chosen from more than 14 thousand stars of the first seven magnitudes, is *not likely to differ much from a certain mean size of them all*. Actually, the size of the stars differ enormously and a mean size is only a purely abstract notion.

Here now is Tutubalin's explanation of February 2011.

Philosophers of science had successfully proved that neither theory nor experiment were of any consequence in science and were not suited for anything. The only possible explanation is that scientific cognition, just like religious cognition, is a miracle and revelation. I provided a hint of theology of science in my paper in Uspekhi Fizicheskikh Nauk vol. 163, No. 7, 1993, pp. 93 – 109.

If you will not colour theologically your investigations, they will not give rise to such interest as they really deserve.

Perhaps most extraordinary events do happen (with an extremely low probability). But suppose that a mathematician had somehow divined the yet unknown Pythagorean proposition. Even then he still has to justify it. At first, he can draw a right triangle, measure its sides etc, then rigorously consider his task.

After reading Tutubalin's paper mentioned above, I am still unable to say anything else on this subject, but I saw a significant statement on p. 98: for two hundred years no progress was made about *the fundamental problem: when does statistical stability emerge*?

I have now found a highly relevant statement by Kolmogorov in the Russian translation of 1986 of his Logical foundations of probability

(*Lect. Notes Math.*, No. 1021, 1983, pp. 1 – 5): Randomness in the wide sense indicates phenomena which do not exhibit regularities, do not necessarily obey any stochastic laws. It should be distinguished from stochastic randomness, a subject of the theory of probability.

Bibliography

Grekova I. (1976 Russian), Peculiar methodological features of applied mathematics on the current stage of its development. *Voprosy Filosofii*, No. 6, pp. 104 – 114.

Fourier J. B. J. (1829), Historical Eloge of the Marquis De Laplace. *London, Endinb. and Dublin Phil. Mag*., ser. 2, vol. 6, pp. 370 – 381. The original French text was only published in 1831.

Herschel W. (1817), Astronomical observations and experiments tending to investigate the local arrangement of celestial bodies in space. *Scient. Papers*, vol. 2. London, 1912, pp. 575 – 591. Reprint of book: London, 2003.

Poincaré H. (1896), *Calcul des probabilités*. Paris, 1912; reprinted 1923.

I

V. N. Tutubalin

Theory of Probability in Natural Science

Teoria Veroiatnostei v Estestvoznanii. Moscow, 1972

Introduction

Even from the time of Laplace, Gauss and Poisson the theory of probability is using a complicated mathematical arsenal. At present, it is applying practically the entire mathematical analysis including the theory of partial differential equations and in addition, beginning with Kolmogorov's classic (1933), measure theory and functional analysis. Nevertheless, books on the theory of probability for a wide circle of readers usually begin by stating that the fundamental problems of applying it are quite simple for a layman to understand. That was Cournot's (1843) opinion, and we wish to repeat his statement right here.

However, it could have been also stated that those problems are difficult even for specialists since scientifically they are still not quite clear. More precisely, when discussing fundamental stochastic problems, a specialist fully mastering its mathematical tools has no advantage over a layman since they do not help here. In this case, important is an experience of concrete applications which for a mathematician is not easier (if not more difficult) to acquire than for an engineer or researcher engaged in direct applications.

At present, ideas about the scope of the theory of probability took shape a bit more perfectly than in the time of Laplace and Cournot. We begin by describing them.

1. Does Each Event Have Probability?

1.1. The concept of statistical stability (of a statistical ensemble). Textbooks on the theory of probability, especially old ones, usually state that each random event has probability whereas a random event is such that can either occur or not. Several examples are offered, such as the occurrence of heads in a coin toss or of rain this evening or a successful passing of an examination by a student etc. As a result, the reader gets an impression that, if we do not know whether a given event happens or not, we may discuss its probability, and the theory of probability thus becomes a science of sciences, or at least an absolutely special science in which some substantial inferences may be reached out of complete ignorance.

Modern science naturally vigorously rejects that understanding of the concept of probability. In general, science prefers experiments whose results are stable, i. e. such that the studied event invariably occurs or not. However, such complete stability of results is not always achievable. Thus, according to the views nowadays accepted in physics, it is impossible for experiments pertaining to quantum mechanics. On the contrary, it can be considered established

sufficiently securely that a careful and honest experimentalist can in many cases achieve statistical, if not complete stability of his results.

As it is now thought, events, connected with such experiments, are indeed comprising the scope of the theory of probability. And so, the possibility of applying the theory of probability is not, generally speaking, presented for free, it is a prize for extensive and painstaking technical and theoretic work on stabilizing the conditions, and therefore the results, of an experiment. But what exactly is meant by statistical stability for which, as just stated, we ought to strive? How to determine whether we have already achieved that desired situation, or should we still perfect something?

It should be recognized that nowadays we do not have an exhaustive answer. Mises (1928/1930) had formulated some pertinent demands. Let μ_A be the number of occurrences of event A in n experiments, then μ_A/n is called the frequency of A. The first demand consisted in that the frequency ought to become near to some number $P(A)$ which is called the probability of the event A and Mises wrote it down as

$$\lim \mu_A/n = P(A),\ n \to \infty.$$

In such a form that demand can not be experimentally checked since it is practically impossible to compel n to tend to infinity.

The second demand consisted in that, if we had agreed beforehand that not all, but only a part of the trials will be considered (for example, trials of even numbers), the frequency of A, calculated accordingly, should be close to the same number $P(A)$; it is certainly presumed that the number of trials is sufficiently large.

Let us begin with the merit of the Mises formulation. Properly speaking, it consists in that some cases in which the application of the theory of probability would have been mistaken, are excluded, and here the second demand is especially typical; the first one is apparently well realized by all those applying the theory of probability and no mistakes are occurring here.

Consider, for example, is it possible to discuss the probability of an article manufactured by a certain shop being defective[1]. One of the causes of defects can be the not quite satisfactory condition of a part of workers, especially after a festive occasion. According to the second Mises demand, we ought to compare the frequency of defective articles manufactured during Mondays and the other days of the week, and the same applies to the end of a quarter, or year due to the rush work. If these frequencies are noticeably different, it is useless to discuss the probability of defective articles. Finally, defective articles can appear because of possible low quality of raw materials, deviation from accepted technology, etc.

Thus, knowing next to nothing about the theory of probability, and only making use of the Mises rules, we see that for applying the theory for analyzing the quality of manufactured articles it is necessary to create beforehand sufficiently adjusted conditions. The theory of probability is something like butter for the porridge: first, you ought to prepare the porridge. However, it should be noted at once that the theory of probability is often most advantageous not when it can be

applied, but when, after attempting to make use of it, a lack of statistical homogeneity (which is the same as stability) is revealed.

If the articles manufactured by a certain shop may be considered as a statistically homogeneous totality, the serious question still is, whether the quality of those articles can be improved without fundamentally perfecting technology. If, however, the quality is fluctuating (which should be stochastically established), then the pertinent cause can undoubtedly be revealed and the quality improved.

The main shortcoming of the Mises formulation is its indefiniteness. It is not stated how large should the number of experiments n be for ensuring the given beforehand closeness of μ_A/n to $P(A)$. A quite satisfactory answer can only be given (see below) after additionally presuming an independence of the results of individual trials. An experimental check of independence is partially possible, but difficult and always, without exception!, incomplete.

But the situation with the Mises second demand is much worse. As formulated above, it is simply contradictory since, indicating beforehand some part of the n trials, we could have accidentally chosen those in which the event A had occurred (or not) and its frequency will be very different from the frequency calculated for all the trials. Mises certainly thought not about selecting any part of the trials, but rather of formulating a reasonable rule for achieving that.

Such a rule should depend on our ideas about the possible ways of corrupting statistical homogeneity. Thus, fearing the consequences of a Sunday drinking bout, we ought to isolate the part of the production manufactured on Mondays; wishing to check the independence of event A from another event B, we form two parts of the trials, one in which B occurred, the other one, when it failed. These reasonable considerations are difficult to apply in the general case, i. e., they can hardly be formulated in the boundaries of a mathematical theory.

We see that there does not exist any mathematically rigorous general method for deciding whether a given event has probability or not. This certainly does not mean that in a particular case we can not be completely sure that stochastic methods may be applied. For example, there can not be even a slightest doubt in that the Brownian motion can be stochastically described. Brownian motion is a disorderly motion of small particles suspended in a liquid and is caused by the shocks of its moving molecules. Here, our certainty is justified rather by general ideas about the kinetic molecular theory than by experimental checks of statistical stability.

In other cases, such as coin tossing, we base our knowledge on the experience of a countless number of gamblers playing heads or tails. Note, however, that many eminent scientists did not think that the equal probability of either outcome was evident. Mises, for example, declared that before experimenting we did not know about it at all; anyway, there is no unique method for deciding about the existence of statistical stability, or, as the physicists say, of a statistical ensemble.

The stochastic approach is therefore never mathematically rigorous (provided that a statistical ensemble does exist) but, anyway, it is not less rigorous than the application of any other mathematical method in natural science. For being convinced, it is sufficient to read § 1 (What

is energy?) from chapter 4 of Feynman (1963). In an excellent style but, regrettably in a passage too long for being quoted, it is stated there that the law of conservation of energy can be corroborated in each concrete case by finding out where did energy go, but that modern physics has no general concept of energy. This does not prevent us from being so sure in that law that we make a laughing-stock of anyone telling us that in a certain case the efficiency was greater than 100%. Many conclusions derived by applying stochastic methods to some statistical ensembles are not less certain than the law of conservation of energy.

The circumstances are quite different for applying the theory of probability when there certainly exists no statistical ensemble or its existence is doubtful. In such cases modern science generally denies the possibility of those applications, but temptation is often strong... Let us first consider the reason why.

1.2. The restrictiveness of the concept of statistical ensemble (statistical homogeneity). The reason is that that concept is rather restrictive. Consider the examples cited above: coin tossing, passing an examination, rainfall. The existence of an ensemble is only doubtless in the first of those. The business is much worse in the other two examples. We may discuss the probability of a successful passing of an examination by a randomly chosen student (better, by that student in a randomly chosen institute and discipline and examined by a randomly chosen instructor). Randomly chosen means chosen in an experiment from a statistical ensemble of experiments. Here, however, that ensemble consists of exactly one non-reproducible experiment and we can not consider that probability.

It is possible to discuss the probability of rainfall during a given day, 11 May, say, of a randomly chosen year, but not of its happening in the evening today. In such a case, when considering that probability in the same morning, we ought to allow for all the weather circumstances, and we certainly will not find any other day with them being exactly the same, for example, with the same synoptic chart, at least during the period when meteorological observations have been made.

Many contributions on applying the concept of stochastic process have appeared recently. It should describe ensembles of such experiments whose outcome is not an event, or even a measurement (that is, not a single number), but a function, for example a path of a Brownian motion. We will not discuss the scope of that concept even if the existence of a statistical ensemble is certain but consider the opposite case. Or, we will cite two concrete problems.

The first one concerns manufacturing. We observe the value of some economic indicator, labour productivity, say, during a number of years (months, days) and wish to forecast its values. It is tempting to apply the theory of forecasting stochastic processes. However, our experiment only provides the observed values and is not in principle reproducible, and there is no statistical ensemble.

The other problem is geological. We measured the content of a useful component in some test points of a deposit and wish to determine its mean content, and thus the reserves if the configuration

of the deposit is known. It is tempting to apply here the theory of estimating the mean of a stochastic process, but here also it is unclear what should constitute the ensemble of realizations. If a new realization is understood as similar values at points chosen along another line, it is unclear whether they will possess the same statistical properties, and still less clear if data pertaining to other deposits are chosen.

These examples are sufficiently important for understanding the wish to create such stochastic methods which will not need ensembles. However, modern probability theory has no such methods but only particular means for saving the concept of statistical homogeneity and even they are not at all universally applicable. So how should we regard the application of the theory of probability in such cases?

1.3. Relations between medicine and magic. The problem stated above resembles that of the relations between medicine and magic whose idea I have borrowed from Feynman (1963) but am considering it in more detail. Suppose we discuss the treatment of malaria, and the shaman knows that the Peruvian bark will help whereas shaking a snake above the patient's face is of no use. So he prescribes in essence the same treatment as a physician will. True, the doctor will give quinine instead of the bark, but this is not very important, and, which is the main point, he knows the life cycle of the plasmodium and will correctly prescribe the duration of the treatment.

The physician has therefore more chances of success, but the main difference between medicine and magic consists in the attitudes of the doctor and the shaman in case of failure. The shaman will explain it by the devil's meddling and do nothing more; the doctor, however, will look for the real cause of failure and hope that such knowledge will at least help other patients if not the first one who could have died. The history of science is a history of ever more precise cognition of reality which is indeed restricting the arbitrary intervention of the devil in whose face the shaman feels himself hopeless.

However, we do not succeed in really banishing the devil. Even in mathematics he is able to interfere which is manifested for example in contradictions; most troublesome are those pertaining to the set theory. A grand attempt to expel the devil from mathematics connected with the names of Bertrand Russell, Hilbert, Gödel, and other first-rate mathematicians had been attempted in the first half of the 20^{th} century, and what did emerge?

It occurred that along with the devil it would have been necessary to banish some notions which we do not at all wish to be deprived of, for example the idea of a number continuum. It is impossible, say (without offering the devil a finger instead of which he will snap off your hand), to state that a function continuous on an interval reaches its maximum value. Such excessively radical exorcism (constructive mathematical analysis) was naturally not recognized; we have to tolerate the devil.

True, for the mathematical theory of probability that devil is actually only an imp who inflicts no special harm. However, I recall that once, desiring to apply transfinite induction (a mathematical trick involving something devilish) for proving a theorem, I discovered much to my relief that the process of induction did not actually

demand to apply transfinite numbers but was rather reduced to usual mathematical induction.

In the applied theory of probability the harmless imp turns out a sharp horned devil who favours to corrupt meanly statistical homogeneity. So far as we keep to the concept of ensemble and check that homogeneity by available methods, we are able at least to reveal in time the devilish dirty trick whereas, abandoning it, we wholly surrender ourselves to the devil's rule and ought to be prepared for surprises. Thus, from the point of view of modern probability theory, the boundary between science and magic is defined by the notion of statistical ensemble. It follows that inferences, derived by applying that theory when a statistical ensemble of experiments is lacking, has no scientific certainty.

Unlike the arsenal of magic, the tools of science must be entirely justified. However, when concluding that, for example, the error of a result obtained from a single realization of a stochastic process is situated in the given interval with probability 0.95, we do not know to what does that probability correspond, – to an ensemble of realizations which we ought to conjuncture by issuing from the single observed realization so as to apply the notion of stochastic process?

But all those other realizations are irrelevant and it is very easy to provide examples of faulty inferences made when applying the theory of probability in manufacturing, geology, etc where it is senseless to discuss statistical ensembles. Historically, science emerged from magic but treats it disdainfully and would wish to ignore it. However, we should not wholly yield to that temptation either.

A representative of the constructive direction in mathematics considers the usual mathematical analysis a magic. We should rather distinguish between white and black magic the latter connected with being subjectively unconscionable. At present, we can not ignore honest attempts to apply probability theory when statistical ensembles are lacking. I venture to forecast that something being magic today will become science tomorrow. It would have been unreasonable to keep too strongly to the established concept of statistical homogeneity. However, here I will entirely hold on to that concept since nowadays any other method of obtaining really plausible results is lacking.

1.4. Summary. Thus, while perfection of experimenting is going on in one or another branch of science or technology, a special situation often arises when statistical stability is present but complete stability of the results is impossible to achieve. The former is characterized by stability of the frequencies of the occurrences of the various events connected with the experiment's outcome.

An exhausting check of such stability (statistical homogeneity, statistical ensemble) is impossible, but in many cases the presence of a statistical ensemble is sufficiently certain. According to modern ideas, these cases indeed comprise the field of scientific applications of the probability theory.

And still there exists a readily understood wish to apply it also in other cases in which the results of the experiments are not definite, but the existence of a statistically homogeneous ensemble is impossible. For the time being, such applications belong to magic rather than

science, but, provided subjective honesty, they can not be ignored. In future it will perhaps be possible to make them scientific. As testified by the entire history of science, its origin had occurred by issuing from factual material collected while practising magic.

2. The Foundations of the Mathematical Arsenal of the Theory of Probability

Modern probability is sharply divided into mathematical and applied parts. Mathematical statistics adjoins the former whereas the latter is closely connected with the so-called *applied statistics*. An attempt to *define* those sciences would have led us into such scholastic jungle, that, terror-stricken, we abandon this thought. Here, we wish to adopt some intermediate stand, and we begin with the mathematical theory of probability.

It busies itself with studying the conclusions of the Kolmogorov axiomatics (1933) and has essentially advanced in developing purely mathematical methods. However, it wholly leaves aside the question of which phenomena of the real world does the axiomatic model correspond to well enough, or somewhat worse, or not at all, respectively. It is possible to adduce really far-fetched examples of mistakes made by mathematicians lacking sufficient experience and practical intuition when attempting to work in applications.

However, the axiomatic model is suitable for developing the mathematical arsenal. There, the generally known stochastic concepts and theorems simply become particular cases of the corresponding concepts and theorems of mathematical analysis. In this chapter, we will indeed describe the pertinent subject. The following chapters are devoted to the substantial stochastic theorems.

The reader ought to bear in mind that this booklet is not a textbook, and that here the theory of probability is therefore dealt with briefly and sometimes summarily. Its knowledge is not formally required, and all the concepts necessary for understanding the following chapters are defined, but examples are not sufficiently numerous. Without them, it is impossible to learn how to apply the axiomatic model, and it would be better if the reader is, or intends to be acquainted with the theory of probability by means of any textbook even if it does not keep to the axiomatic approach. From modern textbooks, we especially advise pt 1 [vol. 1] of Feller (1950).

2.1. Discrete space of elementary events. In the simplest case quite sufficient for solving many problems the entire theory of probability consists of one notion, one axiom and one definition. Here they are.

The concept of stochastic space. A stochastic space Ω is any finite or countable set corresponding to whose elements $\omega_1, \omega_2, ..., \omega_n, ...$ non-negative numbers $P(\omega_i) \geq 0$ called their probabilities are attached. *Set* means here the same as *totality*, that is, something consisting of separate elements. A set is called countable if its elements can be numbered 1, 2, ..., n, ...

We will introduce the notation

$$\Omega = \{\omega_1, \omega_2, ..., \omega_n, ...\}, \text{ or } \Omega = \{\omega_i : i = 1, 2, ..., n, ...\}$$

for stating that Ω consists of elements ω_1, ω_2, ..., ω_n, ... Elements ω_i are also called elementary events or outcomes.

Axiom. The sum of the probabilities of all the elementary events is 1:

$$P(\omega_1) + ... + P(\omega_n) + ... = \sum_{i=1}^{\infty} P(\omega_i) = \sum_{\omega_i \subseteq \Omega} P(\omega_i) = 1.$$

Definition. An event is any subset (part of set) of the set of elementary events; the probability of an event is the sum of probabilities of its elementary events. That set A is a subset of set Ω (i. e., that A consists of some elements included in Ω) is written as $A \subseteq \Omega$. The probability of event A is denoted by $P(A)$ and the definition is written down as

$$P(A) = \sum_{\omega_i \subseteq A} P(\omega_i).$$

The explanation below the symbol of summing means that those and only those $P(\omega_i)$ are summed which are included in A.

The described mathematical model can be applied for very many stochastic problems. However, all of them are initially formulated not in the terminology of the space of elementary events, i. e., not in the axiomatic language but in ordinary terms. This [?] is unavoidable because only by considering problems any student of probability becomes acquainted with those concrete situations in which it is applicable. It is impossible to describe such situations in the axiomatic language and it is therefore necessary to learn how to translate the conditions of problems into the language of elementary events.

The situation here is quite similar to that which school students encounter when solving problems in compiling systems of equations: there, a translation from one language into another one is also needed. Such translations can be either very easy or difficult or ambiguous with differing systems of equations appearing in the same problem. In this last-mentioned case, one such system can be difficult to compile but easy to solve with the alternative system being opposite in that sense (easy and difficult respectively).

We stress therefore that, introducing a space of elementary events corresponding to a given problem, is not a purely mathematical operation as a proof of a theorem, but indeed a translation from one language into another one, and it is senseless to strive for such a rigour as adopted in mathematics. Clear-cut mathematical formulations are now concluded here and we are turning to the rules of translation.

Stochastic problems usually have to do with some experiments, with the set Ω consisting of all its possible outcomes. Thus, in coin tossing Ω consists of two elementary outcomes

$$\Omega = \{\text{heads; tails}\}$$

and in throwing a die there are six such outcomes

$\Omega = \{1, 2, 3, 4, 5, 6\}$.

For the case of two dice Ω consists of all pairs (m, n), showing the numbers of points on them:

$\Omega = [(m, n): m = 1, ..., 6; n = 1, ..., 6]$.

An event A can here be, say, an even sum of the points:

$A = [(m, n): m = 1, ..., 6; n = 1, ..., 6; m + n$ is even$\}$.

In general, descriptions of the set of elementary outcomes are usually easily made, but the situation is quite different when determining the probabilities $P(\omega_i)$ of separate elementary events given the conditions of a problem. According to the frequentist concept of probability it will be necessary to make a large number of experiments and assume the frequencies of the occurrence of the elementary outcomes ω_i as the approximate values of $P(\omega_i)$. This, however, is not always possible; actually, such determinations of probabilities are complicated so that a large part is played by cases in which probabilities can be determined by some speculations without experimenting.

For example, the set Ω rather often consists of a finite number N of elements whose probabilities appear undoubtedly equal to one another. According to the axiom, the probability of each elementary event will then be $1/N$, and if A consists of M elementary events,

$$P(A) = \sum_{\omega_i \subseteq A} P(\omega_i) = \frac{M}{N}. \qquad (2.1)$$

In words: the probability of an event is equal to the ratio of the number of favourable outcomes (outcomes included in the event) to the number of all possible outcomes. When formula (2.1) is applicable, we are discussing a problem in classical probability.

According to modern interpretation, formula (2.1) is not a definition of probability, it is only applicable when all the elementary events are equally probable. And when does this happen? is a rather subtle question. For example, long experiments with dice indicate that their various faces are not generally equally probable; it is difficult to manufacture a perfectly symmetric dice. On the other hand, special measures undertaken when drawing lottery tickets by chance ensure equal probability of winning for each.

To illustrate the possibilities of the mathematical model *we will consider the casting of lots* assuming that such measures were sufficient for ensuring the application of the concept of classical probability. When distributing apartments in a house being built by a cooperative, the casting of lots is sometimes achieved in two stages. At first, lot only decides the order of drawing lots by the members at the second stage, when the actual distribution by chance follows. Is such procedure consisting of two stages necessary? Or, who has more

chances to draw a more suitable apartment, the first or the last in the order of the final drawings?

Suppose there are N apartments, numbers 1, 2, .., n of them worse, and the rest numbers, $n + 1$, ..., N, better. Determine the probability that the member of the cooperative k-th in the order of drawing will draw a worse apartment. The experiment, or drawing the N tickets has outcomes

$i_1, i_2, ..., i_N$, all i_k are different.

Here, i_1 is the number of the apartment drawn by the first person, i_2, same by the second person, etc. The total number of all the possibilities is

$$N(N-1)...2 \cdot 1 = N!.$$

If the tickets are thoroughly shuffled, all the elementary events should be equally probable and we will have a problem in classical probability. Let A_k be the event of the k-th member of the cooperative to draw a worse apartment. In other words, A_k consists of such elementary events $i_1, i_2, ..., i_N$, that i_k takes one of the values 1, 2, ..., n with $i_1, ..., i_{k-1}, i_{k+1}, ..., i_N$ being arbitrary. Let us count the number of those elementary events.

For i_k there are n possibilities;

for $i_1, i_2, ..., i_{k-1}$, there are $N - 1, N - 2, ..., N - k + 1$ possibilities;

for $i_{k+1}, ..., i_N$, there are $(N - k), ..., N - (N - 1) = 1$ possibilities.

Multiplying all the possibilities we see that event A_k consists of $(N - 1)!n = (N!/N)n$ elementary events so that

$$P(A_k) = \frac{(N!/N)n}{N!} = \frac{n}{N}$$

and does not depend on k.

In other words, the probability of choosing a worse apartment can not depend on the order of drawing, so that the first drawings are superfluous. However, we assumed that the tickets were thoroughly shuffled; otherwise the chances of the members of the cooperative are not identical and the first drawings will essentially equalize the chances. It is regrettably unknown how exactly should the tickets be shuffled for ensuring equal chances whereas the method of shuffling adopted for drawing lottery tickets is too tiresome. It follows that drawings of lots in two stages can not be held absolutely superfluous[2].

2.2. Conditional probability. The reader acquainted with urn stochastic models had undoubtedly noted that the model of the space of elementary events is quite isomorphic to the model of extracting balls from an urn and only differs in that different elementary events can now have differing probabilities and the number of these events can be infinite. Indeed, the real part played by the Kolmogorov axiomatics only becomes clear when considering uncountable spaces of elementary events, but even in the simplest case (finite or countable

number of events) the advantage of the axiomatic approach is that it distinctly separates the solution of stochastic problems into two parts:

1. Choice of the mathematical model of the phenomenon or experiment.

2. Calculation within its limits.

We are thus following Descartes' advice: separate each problem into so many parts that they become solvable. The first part, that is, the choice of the mathematical model, is undoubtedly more difficult, and the difficulty, as stated above, lies in determining the probabilities of the elementary events. A formulation of more or less general rules for overcoming this difficulty demands an introduction of some new concepts. We have considered the concept of classical probability; another useful concept is that of conditional probability, but it is expedient to begin by considering usual operations on events. In the set-theoretic context now adopted these operations coincide with those in the set theory.

A sum (*unification*) *of events*. A sum $A \cup B$ of events A and B is an event consisting of those elementary events that enter into A or B (or both).

A product (*intersection*) *of events.* A product AB of events A and B is an event consisting of those elementary events that enter both A and B.

A complementary (*contrary*) *event* $\bar{A}$ of event A is an event composed of those elements that do not enter event A.

If an experiment concludes by one of those elementary outcomes which enter some event C, we say that event C had occurred. Thus, the sum of events A and B occurs if at least one of those events has occurred. The product AB occurs if both events A and B has occurred. Complement $\bar{A}$ of event A occurs if A has not occurred.

Mathematically, conditional probability $P(A/B)$ that A occurs if B has occurred is determined by the equality

$$P(A/B) = \frac{P(AB)}{P(B)},\ P(B) \neq 0.$$

It follows that $P(AB) = P(B)\ P(A/B)$.

The part played by the concept of conditional probability is revealed by its frequentist interpretation. Consider n experiments with events A and B occurring or not in each and let μ_A, μ_B, and μ_{AB} be the number of occurrences of events A, B, and AB. It is evident that μ_{AB} is also the number of occurrences of event A in those experiments, and the ratio μ_{AB}/μ_B –the conditional frequency of event A if event B has occurred. Then

$$\frac{\mu_{AB}}{\mu_B} = \frac{\mu_{AB}/n}{\mu_B/n} \approx \frac{P(AB)}{P(B)} = P(A/B)$$

and the conditional probability is interpreted as the conditional frequency.

Let the space of elementary events Ω be separated into parts B_1, B_2, ..., B_n, so that $\Omega = B_1 \cup B_2 \cup ... \cup B_n$ and no two sets B_i and B_j have common elements. Then, for any $A \subseteq \Omega$ we will have

$$A = AB_1 \cup AB_2 \cup ... \cup AB_n$$

which means that the elementary events included in A are separated into those entering B_1, B_2, ..., B_n and obviously

$$P(A) = P(AB_1) + P(AB_2) + ... + P(AB_n).$$

This follows from the definition of $P(A)$, see § 2.1.

By definition of conditional probability

$$P(A) = \sum_{i=1}^{n} P(AB_i) = \sum_{i=1}^{n} P(B_i)P(AB_i) \qquad (2.2)$$

which is the formula of complete probability.

There is another, the so-called Bayes formula

$$P(B_i / A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A / B_i)}{\sum_{i=1}^{n} P(B_i)P(A / B_i)}. \qquad (2.3)$$

We have derived formulas (2.2) and (2.3) by issuing from the definition of conditional probability and applying really trivial transformations. They can not therefore be called substantial mathematical theorems, but they nevertheless play an important part.

Let us first consider the application of formula (2.2). Suppose, for the sake of definiteness, that event A means that some article is defective and assume also that that event is not by itself statistically stable; more definitely, that there are mutually exclusive conditions of manufacturing B_1, B_2, ..., B_n such that given B_i, it is possible to consider $P(A/B_i)$ so that statistical stability is present.

Suppose now that all the products manufactured under those conditions are stored without being sorted out but that their share corresponding to condition B_i is given and equal to $P(B_i)$. Consider now an experiment in which one article is chosen at random and checked. Two outcomes are possible: A (defective) and $\bar{A}$ (quality sufficient). Its random extraction means that such an experiment is statistically stable, $P(A)$ is expressed by formula (2.2) and

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

Unjustified hope had been previously connected with the Bayes formula (2.3) since subjective interpretation of probability was not ruled out. For example, when having hypotheses B_1, B_2, ..., B_n trusted with probabilities $P(B_1)$, $P(B_2)$, ..., $P(B_n)$, it was thought that an experiment was desirable for indicating the proper hypothesis.

Suppose that event A will occur in that experiment with probability $P(A/B_i)$ if hypothesis B_i is indeed correct. After calculating $P(B_i/A)$ according to that formula, we will obtain new estimates of the likelihood of the various hypotheses.

Modern probability theory considers subjective probability as a concept of magic[3] and only the terminology is preserved according to which probabilities $P(B_1)$, $P(B_2)$, ..., $P(B_n)$ are called prior, and $P(B_1/A)$, $P(B_2/A)$, ..., $P(B_n/A)$, posterior. But magic should be treated carefully: there exists an important scientific domain where the mentioned magical consideration is revived in an undoubtedly scientific manner, the domain of machine diagnostics.

Suppose that a certain hospital admits patients suffering from diseases B_1, B_2, ..., B_n. The prior probabilities $P(B_1)$, $P(B_2)$, ..., $P(B_n)$ are interpreted as frequencies of the corresponding diseases. Event A should be understood here as the totality of the results of a diagnostic examination of a patient. Posterior probabilities $P(B_1/A)$, $P(B_2/A)$, ..., $P(B_n/A)$ offer some objective method of summing the information contained in those examinations; objective does not necessarily mean good enough, but, anyway, not to be neglected beforehand.

The problem is only to find the probabilities $P(A/B_i)$ needed for calculating those posterior probabilities. It seems that for statistically deriving it, it suffices to look at its frequency as given in the case histories of those suffering from B_i, but here we encounter a very unpleasant surprise: A is the result of a large number of examinations, a totality, so to say, of all the indications revealable in a given patient and essential for diagnosing him/her. Even the simplest examination includes nowadays a number of analyses and investigations and partial investigations by many physicians of various specialities. It will not be an exaggeration at all to say that the amount of information is such that 50 binary digits will be needed to write it down; actually, that number will perhaps only suffice after thoroughly selecting the indications essential for the diagnosis.

When adopting these 50, we will have $2^{50} \approx 10^{15}$ various possible values of A. Suppose that previous statistics collected data on 10^4 patients, then, in the mean, 10^{-11} observations will be available for each possible value of A. Practically this means that an overwhelming majority of these values are not covered by any observations, almost each new patient will provide a previously unknown result of examination and it will be absolutely impossible to determine directly the probability $P(A/B_i)$.

Generally speaking, in practical statistical investigations, when desiring to consider at once many factors and connections between them, we usually find ourselves in a blind alley. Classifying statistical material according to several indices very soon provides groups of one observation, and it is not known what to do with them. Then, the Bayes theorem being mathematically trivial naturally can not by itself provide any practical result. Nevertheless, consideration of many factors in medicine is possible. There are contributions whose results are difficult to doubt, but it is premature to describe them for the general reader. One of the possibilities here is connected with applying the concept of independence whose formulation we will now provide.

2.3. Independence. When desiring to consider the complete stochastic characteristic of events $A_1, A_2, ..., A_n$, we will need to know the probabilities of every possible set

$$P(C_1, C_2, ..., C_n)$$

where each C_i can take two values, A_i and $\overline{A}_i$. It is not difficult to calculate that 2^n probabilities are needed. This number increases very rapidly with n and the pertinent possibilities of any experiment become insufficient. We expect such stochastic models to be applicable only if that difficulty is somehow overcome and the main part is played here by the concept of independence.

Definition. Two events, A and B, are independent if the conditional $P(A/B)$ and unconditional probabilities coincide:

$$P(A / B) = \frac{P(AB)}{P(B)} = P(A) \text{ or } P(AB) = P(A)P(B).$$

For n events $A_1, A_2, ..., A_n$ independence is defined by equality

$$P(C_1\, C_2, ..., C_n) = P(C_1)\, P(C_2) \ldots P(C_n) \qquad (2.4)$$

where each C_i can take values A_i and $\overline{A}_i$. Since $P(\overline{A}_i) = 1 - P(A_i)$, the probabilities for independent events can be given by only n values $P(A_1), P(A_2), ..., P(A_n)$.

Independent events do exist; they are realized in experiments carried out independently one from another (in the usual physical meaning). A textbook on the theory of probability should show the reader how the corresponding space of elementary events is constructed here, but this booklet is not a textbook. I have provided a sufficiently detailed exposition of the most essential notions of that theory so as to show how it is done, briefly and conveniently (one concept, one axiom, one definition) in the set-theoretic language. The further development is also offered briefly and conveniently, but from the textbook style I am turning to the style of a summary.

2.4. Random variables. *Definition*. A random variable is a function defined on a set of elementary events. They are usually denoted by Greek letters ξ, η, ζ etc. When desiring to include the argument $\omega \subset \Omega$, we write $\xi(\omega)$, $\eta(\omega)$, $\zeta(\omega)$ etc.

A set of possible values $a_1, a_2, ..., a_n, ...$ of events, all of them different,

$$\{\omega:\ \omega \subset \Omega,\ \xi(\omega) = a_i\} = \{\xi = a_i\}$$

is connected with each random variable $\xi = \xi(\omega)$, as well as probabilities

$$P\{\xi = a_i\} = \sum P(\omega) = p_i,\ \omega{:}\xi(\omega) = a_i.$$

The table

$a_1 \quad a_2 \quad \dots a_n \dots$
$p_1 \quad p_2 \quad \dots p_n \dots$

is called the distribution of the variable ξ.

It should be clearly imagined that, practically speaking, almost always we have to deal not with random variables themselves but only with their distributions. In a word, the reason is that the random variables, being functions of elementary events, are usually unobservable. As a result of an experiment whose outcome is one of the elementary events ω, we usually determine a value of a random variable $\xi(\omega)$, but we will not find out ω.

Let us consider a throw of a die although introducing the set of elementary events in a complicated way understanding ω as the set of values of the coordinates and velocities of the die at the moment when we let it go. More precisely, ω will be the set of those numbers written down precisely enough for uniquely determining the outcome $\xi(\omega)$. Such a determination is not now possible for the microcosm but in our case we do not doubt it although no one ever checked that possibility. In any case, it is extremely difficult to observe ω so precisely, and practically although not in principle even impossible but the observation of $\xi(\omega)$ is easy, and that is what the gamblers are only doing. The space of elementary events Ω is extremely convenient as a concept, as we have seen and will see in the sequel, but as a rule it is not actually observable. It is easier to observe events of the kind $\{\xi = a_i\}$.

And still, such events are too numerous and it is preferable to characterize the distribution of a random variable by several parameters, i. e. by functions of the values a_i and probabilities p_i. Considered are not arbitrary distributions, but such as are uniquely determined by a small number of parameters. Fine, if one or two parameters is (are) needed, endurable if three or four. However, determine experimentally more than four parameters, and your results will be questioned. The point is, that, as empirically noted, when selecting too many parameters any experimental results can be fitted to any law of distribution.

Expectation is the most important parameter of distribution. We will define it not in its usual form; the generally accepted definition will appear as a very simple theorem.

Definition. An expectation of a random variable $\xi = \xi(\omega)$ is number $E\xi$ determined by the formula

$$E\xi = \sum \xi(\omega) P(\omega), \omega \in \Omega.$$

It is assumed here that the series absolutely converges; otherwise, the random variable is said to have no expectation.

It is not difficult to convince ourselves that our definition actually coincides with the accepted formula [...]

$$\mathrm{E}\xi = \sum a_i p_i .$$

Our form of definition is however more convenient for proving the theorems on the properties of the expectation. Let us prove, for example, [the theorem about the expectation of a sum of variables]. [...] In many textbooks that statement is proved defectively. [...]

The second most important parameter of the distribution of a random variable is variance.

Definition. [...]

For random variables as also for events, the concept of independence is most important. We define independence (in totality) for three random variables, and the definition is similar for any number of them.

Definition. [...]

We will prove that for independent random variables

$$\mathrm{E}(\xi \cdot \eta \cdot \varsigma) = \mathrm{E}\xi \cdot \mathrm{E}\eta \cdot \mathrm{E}\varsigma .$$

Proof. [...]

It easily follows that the variance of a sum of independent random variables is equal to the sum of the variances of its terms.

We have concluded the exposition of the main stochastic concepts for the discrete case, when the experiment has [only] a finite or a countable number of elementary outcomes. Now, we have to consider what happens when it is more natural to describe the experiment by a more complicated space.

2.5. Transition to the general space of elementary events. If an experiment results in some measurement, it is possible to state that, since the precision of all measurements is only finite, the set of elementary outcomes will at most be countable. However, the history of the development of science indicates that physical theories are much simplified by considering continuous models for which experimental results can be any number. Differential equations can only be applied in such models.

Readers, familiar with difference equations will easily imagine how more elegant and simple are the differential equations. Thus, although modern physics has some vague ideas about the possible discreteness of space, it certainly is not at all easy to abandon the notion of continuum. And, allowing that notion, what kind of probability theory should we have? The answer to this question is given by the celebrated Kolmogorov axiomatics (Kolmogorov 1933; Feller 1950 and 1966). Its foundation is the notion of the space of elementary outcomes Ω which can now be arbitrary. Some (but not all!) of its subspaces are held to be so to say observable as an experimental result and called events. If A is an event, we are able to say whether it occurred in an experiment or not and in this sense it is observable. We may thus discuss the frequency of its occurrence and consequently the probability $P(A)$.

The main demand of the Kolmogorov axiomatics containing as though in embryo the merits and shortcomings of the entire theory is that, given a countable set of events $A_1, A_2, ..., A_n, ...$, their sum and

intersection are also events; in addition, it is also assumed that Ω is an event with $P(\Omega) = 1$ and that the complement of any event is also an event.

Concerning probabilities, the following fundamental property is assumed. If the events $A_1, A_2, ..., A_n, ...$ do not intersect in pairs (have no common events)

$$P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P(A_i), \tag{2.5}$$

where the symbol U means a sum. For the discrete case, this statement can be declared a theorem derived by issuing from the mentioned definition of § 3.1. In the general case, it is an axiom whereas that definition is useless.

We will consider what does the application of the Kolmogorov axiomatics demand by discussing a concrete example, experimental random throws of a point on interval [0, 1]. Here, the space of elementary outcomes Ω should apparently consist of all points of that interval. If $0 \le a < b \le 1$, it would have been extremely annoying to be forbidden to discuss the probability of a random point ω occurring within interval $[a, b]$. And so, we desire to call events sets of the kind

$$\{\omega: a \le \omega \le b\}$$

and we will assume that

$$P\{\omega: a \le \omega \le b\} = b - a,$$

or, that the probability of a random point to fall on an interval is equal to the interval's length. So far, everything is natural.

Now, however, we must assume that events are not only intervals, but anything obtainable from them by summing and intersecting their countable number as also by including complements. Selecting point c, $0 \le c \le 1$, and a sequence of intervals $[c - 1/n, c + 1/n]$, we see that the intersection of their countable number consists of a single point c, so that any point is an event. The set of rational points is obtained by summing a countable number of points and is therefore an event. The set of irrational points is its complement and therefore also an event.

We thus consider observable whether a point thrown on an interval is rational or irrational although physically this is impossible, and we see that it is necessary to apply carefully the Kolmogorov model, otherwise it can lead to physically absurd corollaries.

Particularly complicated versions of such models are applied in the theory of stochastic processes. There, the researcher ought to be especially careful, ought to possess a certain taste for natural science. Otherwise it is easy to derive such results by issuing from the accepted mathematical model which at best can not be physically interpreted, and at worst offer an occasion for a wrong interpretation. As an example, I cite a mathematical theorem according to which the coefficient of diffusion of the Brownian motion can be determined

absolutely precisely if the pertinent path during any however short interval of time is known.

You can encounter a viewpoint stating that a practical estimate of the coefficient of diffusion does not therefore present any difficulties. This opinion has been established to some extent in the literature on the statistics of stationary processes, but it is completely wrong. Two circumstances prevent its application to real Brownian motion. First, the mathematical *Brownian motion*, i. e., the Wiener process, does not describe the real process over short intervals of time whereas exactly the change of the position of the *particle* during infinitely short intervals enters the estimation of the coefficient of diffusion. Second, the idea of knowing exactly the path of some stochastic process during some interval of time is absolutely unrealistic; we do not at all know how to define precisely a non-regularly changing function which is not describable by an analytic expression. I am unable to dwell here in more detail on the theory of stochastic processes and am returning to probability *P*. For intervals, it coincides with their length.

However, it is possible to construct very complicated sets of intervals and mathematical correctness demands that it be possible to define additionally that probability for all such sets while retaining the main property of countable additivity (2.5). The French mathematician Lebesgue provided a construction (the Lebesgue measure) allowing to ascertain the possibility of such an additional definition. It is complicated and we will not discuss it here. However, it can be applied for spaces Ω of a very general kind, consisting for example of functions which is important for the theory of stochastic processes.

Until now, we have discussed the complications necessarily demanded by the Kolmogorov axiomatics; on the other hand, it is however connected with most important simplifications. The introduction of a measure having the property of countable additivity allows to apply the concept of Lebesgue integral; as a concept, it is incomparably simpler and more general than the Riemann integral. In the general case, all the main notions of the theory of random variables occur not more complicated than those described above for the discrete case. Thus, a remarkable simplicity, generality and order is originated in the main notions of the theory of probability. However, the Lebesgue integral is not more than a concept. No one calculates integrals by applying the Lebesgue extension of measure, the Riemann integral is preferred.

It is necessary to mention here a certain difficulty that takes place when teaching mathematical analysis, both at home and abroad. In general, nothing negative can be said about its part dealing with functions of one variable, although it is somewhat tedious; the horror begins with the transition to functions of several variables. The treatment of the differential, and especially integral calculus is here nowadays absolutely unsatisfactory. Take for example the set of the Green, Stokes and Ostrogradsky formulas introduced without any connection between them. Indeed, there exists now a united viewpoint about all of them and it even includes the Newton – Leibniz formula. It is not treated in textbooks, but can be read in Arnold's lectures (1968) on theoretical mechanics.

The exposition of the theory of probability also suffers from that circumstance although less than theoretical mechanics. We are therefore unable to apply either the notion of the Lebesgue integral or a number of useful properties of the ordinary multiple integral and are restricting the description to a necessary minimum. Just as in the discrete case, we pass on from random variables themselves to their distributions, but our deliberations ought to be suitable for several variables at once rather than for one only. In other words, we will consider vector $\xi = \xi(\xi_1, \xi_2, ..., \xi_n)$. Our main principle is to introduce such characteristics that admit an easy transition from one coordinate system to another one although a so-called joint distribution function

$$F_{\xi_1,\xi_2,...,\xi_n}(x_1, x_2, ..., x_n) = P(\xi_1 < x_1, \xi_2 < x_2, ..., \xi_n < x_n)$$

has been applied instead. The transition from coordinates $x_1, x_2, ..., x_n$ to other coordinates $y_1, y_2, ..., y_n$ becomes not only difficult, it is even impossible to describe that procedure by a formula without actually introducing a stochastic measure

$$\mu_\xi(A) = \mu_{\xi_1,\xi_2,...,\xi_n}(A) = P\{\xi = (\xi_1, \xi_2, ..., \xi_n) \in A\}.$$

Here, the vector $\xi = \xi(\xi_1, \xi_2, ..., \xi_n)$ is an event, an element of the set A. The joint distribution function is thus practically useless. Actually, we have to apply density

$$p_\xi(x) = p_{\xi_1,\xi_2,...,\xi_n}(x_1, ..., x_n).$$

It is defined by demanding that for any (not too complicated) set A in a many-dimensional space

$$P\{\xi \in A\} = \int ... \int p_{\xi_1,\xi_2,...,\xi_n}(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n.$$

The integration is over set A. Density plays here the same part as distribution of a random variable in the discrete case. In particular,

$$\mathrm{E} f(\xi_1, ..., \xi_n) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, ..., x_n) p_{\xi_1,...,\xi_n}(x_1, ..., x_n) dx_1 ... dx_n.$$

Most important is the formula connecting the densities of a random vector in various systems of coordinates, a particular case of the formula for the change of the variables in multiple integrals, and I do not introduce it. Note that usual courses in mathematical analysis even lack the necessary notation.

The densities of distribution of the sum, the product, ratio and other operations on random variables can be immediately derived by issuing from it. On the contrary, for one-dimensional variables the notion of distribution function is very useful. Here is its definition:

$$F_\xi(x) = P(\xi < x)$$

where x is any real number. If density of distribution $p_\xi(x)$ exists, then

$$F_\xi(x) = \int_{-\infty}^{x} p_\xi(x)dx.$$

I also introduce the formulas for expectation and variance in this case:

$$\mathrm{E}\xi = \int_{-\infty}^{\infty} xp_\xi(x)dx,\ \mathrm{E}f(\xi) = \int_{-\infty}^{\infty} f(x)p_\xi(x)dx,$$

$$\mathrm{var}\,\xi = \mathrm{E}(\xi - \mathrm{E}\xi)^2 = \int_{-\infty}^{\infty} (x - \mathrm{E}\xi)^2 p_\xi(x)dx.$$

3. Bernoulli Trials. The Poisson Jurors

3.1. Bernoulli trials. And so, it is incomparably simpler to introduce probabilities of independent, rather than dependent events. Therefore, stochastic models with independent events have much more chances to be practically applied. The most simple and thus the most widely applicable is the model in which we imagine a certain number n of independent trials, each of them resulting in one of the two possible outcomes called *success* and *failure*. The probability of success is supposed to be the same throughout and is denoted by p so that failure will be $q = 1 - p$. Denote also success and failure by 1 and 0, then the result of n trials will be a sequence of these numbers having length n.

The set of elementary events ω, $\Omega = \{\omega\}$, thus consists of all such sequences of length n and therefore has 2^n elements. Taking independence of individual trials into account, we ought to provide a definition according to which the probability $p(\omega)$ of each elementary event ω will be calculated by changing each 1 by number p, and each failure by changing each 0 by number q and multiply the obtained numbers. We will then have

$$P(\omega) = p^{\mu(\omega)}q^{n-\mu(\omega)}$$

where $\mu(\omega)$ is the number of unities in the sequence of the ω's.

Experiments described by this stochastic models are called Bernoulli trials, and the random variable $\mu = \mu(\omega)$ is the number of successes in n such trials. Let us determine the distribution of that random variable. Its possible values are evidently numbers 0, 1, ..., n so that

$$P\{\mu = m\} = \sum P(\omega) = \sum p^{\mu(\omega)}q^{n-\mu(\omega)} =$$
$$\sum p^m q^{n-m} = p^m q^{n-m} \cdot (\text{number of such } \omega \text{ that } \mu(\omega) = m).$$

The summations are over ω: μ(ω)= *m*. However, the number of such sequences of ω's that μ(ω) = *m* is clearly equal to the number of possible selections of *m* symbols out of *n*, C_n^m. And so,

$$P\{\ \mu = m\} = C_n^m p^m q^{n-m} \qquad (3.1)$$

which is the main formula of the Bernoulli trials.

Its theory is seen to be almost trivial but not trivial is to learn how to apply it, that is, how to find those phenomena that are sufficiently well described by that pattern. A classical example of the trials is a toss of a coin, but when attempting to discover something more interesting, we enter the domain of doubtfulness. Thus, is it possible to consider a birth of an infant of one or another sex as a Bernoulli trial (and regard a male birth, say, as a success)?

According to genetic ideas, this is quite natural. However, those ideas lead just as naturally to the frequency of male births $p = 1/2$ whereas it somewhat exceeds 1/2 as established by examining such an immense material that it becomes impossible to question it. Then, however, it is perhaps permissible to admit the opposite hypothesis of $p \neq 1/2$? Once more, no, since the Bernoulli trials presume a constant probability of success whereas the statistical data certainly indicate that the frequency of male births increases after long wars. The dependence of the probability of male births on the social conditions of the family [and on other circumstances] is also being discussed so that the model of Bernoulli trials does not in this case completely correspond to reality.

Then, statistically investigating that frequency we find out that, strictly speaking, the model of those trials is unacceptable; however, since the probability of male birth is nevertheless very near to 1/2, it is only possible to reject the hypothesis of its applicability through statistical research based on profound corollaries of formula (3.1). We will see now how it is carried out in Chapter 4.

An application of stochastic methods results in a conclusion that, strictly speaking, we ought not to discuss the probability of male births (or statistical stability). However, in the final analysis we will find out much more than had there been an ideal conformity with the theory of probability: we discover for sure that there exists a still unidentified agent regulating the numbers of men and women.

The model of Bernoulli trials is often applied for estimating some plans of acceptance inspection in which the manufacturing of faulty (failures) or suitable (successes) articles must be described by that pattern. However, after recalling the discussion in Chapter 1 of the possibility of a stochastic description of manufacturing faulty products, it becomes evident that that model can only be made use of when the industrial process is arranged well enough.

We will discuss at length the attempt to apply the same model to the problem of legal verdicts. Pertinent investigations are connected with the names of such first-rate scholars as Laplace and Poisson, and their study is very instructive. It shows by an example taken from history that a perfect command of the mathematical methods of the theory of

probability can be coupled with an absolutely wrong approach to reality[4].

3.2. Poisson's jurors. Laplace, and then Poisson investigated the issue of the probabilities of mistaken legal verdicts. A certain juror can naturally make a mistake. Laplace assigned jurors a very modest ability of correct judgement: he thought that for each separately considered juror the probability of a mistake was a random variable uniformly distributed on segment [0, 1/2]. Poisson did not agree; he rather believed that the probability of a correct judgement should be estimated by issuing from statistical data. The impossibility of precisely establishing whether rightly or not a given accused person was found guilty presents here the greatest difficulty of a direct statistical estimate.

Poisson's ideas widely applied now also consisted in that in such a situation it was necessary to construct a statistical model with the unknown probability entering it as a parameter and to attempt to determine it by pertinent data.

Let us consider the administration of justice in more detail. The trial is based on the inquest. Denote the event consisting in that the evidence collected at the inquest was sufficient for the trial to declare the defendant guilty by A, and the contrary event by $\overline{A}$. Given A, all the jurors, provided their judgement is faultless, ought to unanimously vote for the prosecution; otherwise (event $\overline{A}$) for the defence.

Actually, rather often the votes are divided owing to mistakes made by the jurors. Poisson's main proposition was that such division conformed to the Bernoulli pattern. If n is the number of jurors, p, the probability of a correct judgement of each juror, the number of votes for the prosecution, μ, it is described in the following way.

1) Given A, μ is the number of successes for the n pertinent Bernoulli trials with probability of success p.

2) Given $\overline{A}$, μ is the number of failures for the same pattern.

According to the French legislation, $n = 12$ and the defendant was declared guilty if $\mu \geq 7$. The probability of that outcome is

$$P_g = P(A)P\{\mu \geq 7/A\} + P(\overline{A})P\{\mu \geq 7/\overline{A}\} = P(A)\sum_{m=7}^{12} C_{12}^m p^m (1-p)^{12-m} + [1-P(A)]\sum_{m=7}^{12} C_{12}^m p^{12-m}(1-p)^m. \quad (3.2)$$

Criminal statistics provides the frequency of such verdicts which is approximately equal to P_g and Poisson thoroughly checked its stability over the years. However, expression (3.2) includes two unknown parameters, $P(A)$ and p. Knowing only P_g, it is impossible to determine them and it is therefore necessary to turn to statistics which will indicate not only whether defendants were found guilty or exonerated, but [in one case, see below] by how many votes as well. Thus, being accused exactly by seven votes has probability

$$P_g\{\mu = 7\} = P(A)P\{\mu = 7/A\} + P(\overline{A})P\{\mu = 7/\overline{A}\} = P(A)C_{12}^7 p^7 (1-p)^5 + [1-P(A)]C_{12}^7 p^5 (1-p)^7. \quad (3.3)$$

Knowing the left parts of relations (3.2) and (3.3) approximately equal to the frequencies provided by the criminal statistics it is possible in principle to determine both $P(A)$ and p. Equations (3.2) and (3.3) are of a high degree and their solution is not easy. Poisson, however, developed a general method of their solution and finally successfully solved them. In that, the 19th century, following Laplace and Poisson, problems on probabilities of verdicts entered all textbooks on probability theory, but in the next century such applications of the theory were declared absolutely nonsensical. We ought to find out the reason why.

Poisson's main presumption was independence of the jurors' individual judgements. Fully understanding the need to check the stability of frequencies, he (1837) did not say a word about an experimental check of independence. How was such a procedure possible? When solving equations (3.2) and (3.3), Poisson found out that the probability of a correct judgement of an individual juror approximately equalled 2/3 so that a correct unanimous accusation had probability $(2/3)^{12} < 0.01$ and was almost impossible. However, in neighbouring England, as Poisson himself noted, the law demanded a unanimous decision of all the 12 jurors, and English courts pronounced much more condemning sentences, death sentences included, than the courts in France. To remind, the exposition concerned the 19th century.

Poisson considered that circumstance as a cause for national pride, England was seen as a much less civilized nation although it should have been seen as an argument for doubting his own stochastic model. True, it should be said in all fairness that anyway he was unable to check it given the French criminal statistics. Indeed, protecting the secret of the jurors' voting, the French judicial code did not demand to indicate the number of condemning votes the only exception having been the case of the minimal necessary votes.

Thus, from the modern viewpoint, Poisson's error, formally speaking, consisted in recommending a stochastic model without checking it. He determined two unknown parameters by two observed magnitudes with no possibility of such checking. It is interesting to describe the pertinent opinion of Cournot (1843). Poisson's contemporary, he apparently was not as mathematically powerful as Poisson, much less as Laplace. However, we ought to recognize that he possessed more common sense of a natural scientist, than those first-rate scholars.

In particular, he clearly understood that independence of the jurors' judgement was only a premise that should have been experimentally checked. He even proposed such a change of the judicial code which, without violating the secret of the jurors' voting, would have allowed to obtain the necessary statistical data. As to the independence itself, Cournot believed that, if it did not exist in all the totality of legal proceedings in general, then in any case legislation can be separated into groups of independent cases. He even found out that two such groups concerning crimes against the person and against property will have very near to each other values of the parameters $P(A)$ and p as determined according to the Poisson method.

Nowadays we are sure that no independence of the judgement of individual jurors does exist, so that the groups isolated by Cournot would have most likely consisted of one case only. True, this statement is not really proven so that according to modern science Cournot's point of view is formally invulnerable which once again confirms that he had essentially outstripped his time.

For our days, an important conclusion from the above is that it is by no means permissible to use all the available statistical information for determining the parameters of a statistical model; it is absolutely necessary to leave some part of it for checking the model itself, otherwise, great scientific efforts can result in complete rubbish.

4. Substantial Theorems of the Theory of Probability

4.1. The Poisson theorem. When compiling his treatise, Poisson (1837) discovered one of the main statistical laws. Calculating the probabilities $P(\mu = m)$ that m successes will be achieved in n Bernoulli trials, he found out an approximate formula for large values of n and small values of p:

$$P\{\mu = m\} \approx \frac{\lambda^m}{m!} e^{-\lambda} \qquad (4.1)$$

where $\lambda = np$; for more details see Gnedenko (1950). The exact expression for $P\{\mu = m\}$ depends on three parameters, n, m and p; in the approximate expression, n and p are combined into one.

At first sight this simplification seems trivial and Poisson himself did not think that his formula was really important. Indeed, his treatise included a large number of more precise and almost as suitable formulas. However, the combining mentioned allows to compile a comparatively short table for calculating (4.1) with two entries, m and λ, whereas the precise expression for $P\{\mu = m\}$ would have demanded a table with three entries which is not done yet in a sufficiently convenient form.

Nevertheless, the main role of the formula (4.1) consists not in convenient calculation. Strictly speaking, we express it as a mathematical theorem (Gnedenko 1950) concerning Bernoulli trials, i. e. independent trials with two outcomes and a constant probability of success. The most important circumstance is that those conditions may be violated without denying its conclusion, that is, the equality (4.1).

For example, we may admit that different trials have differing probabilities of success

$$p_1, p_2, \ldots, p_n, \ldots$$

(if only all of them are low). Then the exact expression for $P\{\mu = m\}$ from Chapter 3 as well as good enough approximate expressions become useless (because they are too exact). The comparatively rough expression (4.1) remains valid if only $p_1 + p_2 + \ldots + p_n$, or, if desired, $n\bar{p}$ be substituted instead of λ. It follows that for calculating λ it is not necessary to know the values of p_i, suffice it to know one single parameter, their mean, the new value of the probability of success.

Note also that we often are unable to repeat an experiment with a given probability of success for a sufficiently large number of times so that it is perhaps impossible to find out p_i, which does not necessarily prevent us from calculating $\overline{p}$. The approximate value (4.1) therefore has much more chances to find practical application than the exact formula for $P\{\mu = m\}$.

Then, the demand of independence of the individual trials can be weakened: it may be assumed that they have more than two outcomes but such possibilities exceed the limits of this booklet.

Those possibilities of weakening the conditions of the Poisson theorem without changing its conclusion lead to the Poisson distribution attaching probabilities (4.1) to values $m = 0, 1, 2, ...$ becoming one of the most universal laws. Consider for example the problem of the number of refusals during time T in cases of complicated systems. Suppose a system consists of n elements and $p_i = p_i(T)$ is the probability of a refusal of the i-th element (and that after refusal the damaged element is instantly replaced). The number of refusals is the number of successes in n trials with the i-th trial being connected with the i-th element and its success means a refusal of that element. A given element can experience more than one refusal and its refusal can somewhat influence the refusals of the other elements, – all the same, if p_i are sufficiently low, we expect the Poisson distribution with the parameter $\lambda = n\overline{p}$ to describe the number of refusals.

Deviations are only possible if the connections between the refusals of different elements are strong. Given low values of p_i it is natural to expect a linear dependence of $p_i = p_i(T)$ on T:

$$p_i(T) = p_i' T.$$

Then

$$\lambda = \lambda(T) = \lambda' T \qquad (4.3)$$

and the probability of m refusals will be

$$P\{\mu = m\} \approx \frac{(\lambda' T)^m}{m!} e^{-\lambda' T}.$$

For the probability of failure-free work during time T, that is, for $\mu = 0$, we have

$$P\{\mu = 0\} \approx e^{-\lambda' T}$$

which is the generally known exponent law for the time of failure-free work.

The Poisson and the exponent laws therefore correspond to each other. There occurs some harmonious correspondence that we may hope to apply beneficially for solving practical problems.

Our model does not allow for aging; to achieve that we ought to replace the linear dependence (4.3) by a more complicated dependence

$$\lambda = \lambda(T)$$

with $\lambda(T)$ being actually approximated by a function of a most simple kind, for example by a polynomial. Its coefficients can be obtained by one or another method, for instance by the method of least squares. However, the number of parameters necessary to be included will in this case be larger than when aging is not allowed for and, accordingly, the model will enjoy less faith.

In general, with or without allowing for aging, it is natural to apply the Poisson law for describing the number of failures, only its parameter is determined in differing ways. The fit of the Poisson law, its agreement with the actual data should be checked by statistical tests. If a good agreement is lacking, it will be likely more natural to suspect the statistical homogeneity of the data rather then the applicability of the Poisson law. Only after checking that out may we think about choosing another distribution for describing the number of failures.

None of this certainly applies to the case of strong ties between the failure of different elements. For example, if the failure of one part of the system automatically leads to a failure of its other part, the total number of failures will be doubled. In such cases, even if the number of initial failures follows the Poisson law, the doubled figure will not, and it is more natural to apply here the normal law. And the Poisson law is certainly only applicable when a failure really is a rare event.

An excellent set of other examples of the application of the Poisson law is to be found in Feller (1950) only the theory of rare excursions of stochastic processes can possibly be added to it. Just as any rare event, the number of such excursions beyond a sufficiently high level obeys the Poisson law.

4.2. The Central Limit Theorem. The Poisson law is determined by a single parameter λ. It is not difficult to show that λ is the expectation of a random variable distributed according to that law. Here, we will consider an even more universal stochastic law, the so-called normal law determined by two parameters, expectation and variance. It was discovered at about the same time by Gauss and Laplace who issued from absolutely different considerations. Gauss discovered that exactly in the case of a normal law of distribution of the observational errors it is most natural to choose the arithmetic mean of the individual measurements as the estimator of the real value of the measured magnitude. Laplace's starting point was his discovery of an extremely powerful method of calculating the distribution of a sum of random variables. Gauss' ideas are important for treating the results of measurement and were further developed in mathematical statistics as the so-called method of maximum likelihood. Laplace's ideas concerned the properties of arithmetic operations on a large number of random variables and actually constitute the foundation of modern probability theory. In this booklet, devoted to the theory of probability rather then mathematical statistics, we adopt the Laplacean approach[5].

Consider arbitrary independent random variables

$$\xi_1, \xi_2, ..., \xi_n, ... \qquad (4.4)$$

taking the same values

$$0, \pm 1, \pm 2, ..., \pm m$$

with the same probabilities

$$P\{\xi_i = m\} = p_m.$$

Such variables are called identically distributed. Probabilities p_m are arbitrary, they only obey the condition of adding up to 1 and sufficiently rapidly decrease as $m \to \pm \infty$. More precisely, it is necessary that the variables ξ_i have a finite expectation and a finite variance

$$\mathrm{E}\xi_i = \sum_{m=-\infty}^{\infty} m p_m = a, \ \mathrm{var}\xi_i = \sum_{m=-\infty}^{\infty} (m-a)^2 p_m = \sigma^2.$$

Otherwise, the set of probabilities $\{p_m\}$ is absolutely arbitrary. It can therefore be impossible to describe that set by any finite number of parameters.

Laplace discovered that for a large number of terms of the set (4.4) the distribution of their sum becomes incomparably simpler than that of their separate terms so that, allowing for some additional conditions (Gnedenko 1950, Chapter 8)[6],

$$\sigma\sqrt{n} P\{\xi_1 + ... + \xi_n = m\} \approx \frac{1}{\sqrt{2\pi}} \exp[-\frac{-x_n^2(n)}{2}],$$

$$x_n(m) = \frac{m - na}{\sigma\sqrt{n}}. \qquad (4.5a,b)$$

It is beneficial to bear in mind the following simple considerations which help to understand the geometric meaning of equality (4.5). Suppose that we desire to show graphically the distribution of each random variable (4.4) and their sum. We choose an abscissa axis, indicate points

$$0, \pm 1, \pm 2, ..., \pm m ...$$

and show probability as a rectangle with base 1, its midpoint being at m, and area (that is, its height) p_m. We will have some, generally speaking, irregular set of rectangles. An attempt to show the distribution of the probabilities of the sum of those variables for a large n will be unsuccessful because the possible values of that sum can be very large and the probabilities of the separate values, small, as can be proven. A change of the scale will be therefore needed so that showing the values of the random variable

$(1/B_n)(m - A_n)$

instead of the value of the sum $m = (\xi_1 + ...+ \xi_n)$ will be necessary.

And now the essence of Laplace's discovery can be expressed by a single phrase: the figure should be shifted by

$$A_n = \mathrm{E}(\xi_1 + ... + \xi_n) = na$$

and the coefficient of the change of the scale should be equal to

$$B_n = \sqrt{\mathrm{var}(\xi_1 + ... + \xi_n)} = \sigma\sqrt{n}.$$

The random variable

$$s_n^* = \frac{1}{\sqrt{\mathrm{var}(\xi_1 + ... + \xi_n)}}[(\xi_1 + ... + \xi_n) - \mathrm{E}(\xi_1 + ... + \xi_n)] \qquad (4.6)$$

is called the normed sum. Obviously,

$$\mathrm{E}s_n^* = 0, \quad \mathrm{var}\, s_n^* = 1$$

and the numbers (4.5b) are the possible values of that normed sum. Let us attempt to show its probabilities as rectangles with bases

$$x_n(m+1) - x_n(m) = \frac{1}{\sigma\sqrt{n}},$$

their midpoints coinciding with points $x_n(m)$ and areas equal to probabilities

$$P\{s_n^* = x_n(m)\} = P\{\xi_1 + ... + \xi_n = m\}.$$

The heights of these rectangles should be

$$\sigma\sqrt{n}P\{s_n^* = x_n(m)\} = \sigma\sqrt{n}P\{\xi_1 + ... + \xi_n = m\}.$$

Thus, because of (4.5a) the upper bases of these rectangles will be almost exactly situated along a curve described by equation

$$y = y(x) = \frac{1}{\sqrt{2\pi}}\exp[-\frac{x^2}{2}] \qquad (4.7)$$

independent of anything and calculated once and for all.

A result absolutely not foreseen and almost miraculous! Disorder in probabilities p_m somehow gives birth to a unique curve (4.7) which simply occurs by summing random variables and transforming the scale of the figure. That is Laplace's remarkable discovery without

which the theory of probability would have almost lacked original (not reducible to known concepts of mathematical analysis) contents.

True, the statement (4.7) has some exceptions. For example, if all the random variables (4.4) only take even values (so that $p_m \neq 0$ only for even values of m) the sum $(\xi_1 + ...+ \xi_n)$ is also always even, whereas at odd values of m the left part of (4.5a) vanishes and that equality is violated. This exception is actually the only one (Gnedenko 1950, Chapter 8) and, for avoiding it and because of a number of other considerations, it is preferred to formulate the central limit theorem in terms of distribution functions. Here the appropriate formulation is effectively known to Laplace.

Theorem. Let $\xi_1, ..., \xi_n, ...$ be a sequence of independent identically distributed random variables having a finite expectation a and finite variance σ^2, and suppose that s_n^*, see (4.6), is the normed sum of those variables. Then, as $n \to \infty$,

$$P\{s_n^* < x\} \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp[-\frac{x^2}{2}]dx$$

and for $-\infty < x < \infty$ the convergence is uniform for every x.

This formula is still less sensitive to violations of its conditions then even the Poisson theorem remarkable in this connection. The development of the theory of probability is essentially linked with the perfection of its proof and weakening of its conditions. It is possible to deny, i. e. to replace by less restrictive each of the latter without invalidating its conclusion. Liapunov denied the identical distribution of the random variables and thus occurred his theorem (Gnedenko 1950, Chapter 8) whereas Bernstein (1926) denied independence.

Attempts were recently made to abandon essentially the condition of randomness of the variables (4.4). It is also possible to replace random variables taking numerical values by random elements of some groups (and to consider the relevant group operation instead of summing). Certain success was achieved but it is too soon to discuss this subject here. The finiteness of the variance can not be denied (Gnedenko & Kolmogorov 1949) since the convergence to the normal law will not hold. It is only possible to weaken somewhat that condition.

4.3. The normal distribution. And so, the most widely applied distribution of probabilities is the Gauss – Laplace law whose density is provided by formula (4.7). In other words, it is said that the random variable ξ has a standard normal distribution if (practically for any) set A the equation

$$P\{\xi \in A\} = \int_A \frac{1}{\sqrt{2\pi}} \exp[-\frac{x^2}{2}]dx$$

is valid.

It is natural to consider random variables of the type

$$\eta = \sigma\xi + a$$

along with ξ. For example, if ξ is the result of some measurement, and η is its result in another system of units, it is not difficult to show that the density of distribution of η is

$$p_\eta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-a)^2}{2\sigma^2}]$$

with expectation and variance of η being Eη =*a* and varη = σ^2. The distribution of the random variable η is called normal with parameters *a* and σ and denoted by N(*a*, σ).

Most important is the following theorem: If η_1 and η_2 are independent random variables having distributions N(a_1, σ_1) and N(a_2, σ_2), their sum will have distribution N($a_1 + a_2$, $\sqrt{\sigma_1^2 + \sigma_2^2}$).

4.4. The De Moivre – Laplace theorem. Consider *n* Bernoulli trials with probability of success *p* in each. The number of successes μ can be represented as a sum

$$\mu = \mu_1 + \mu_2 + ... + \mu_n$$

where the random variable μ_k is the number of successes in the *k*-th trial, i. e., is equal to 1 or 0 for achieving a success or not. Since all trials are identical, μ is the sum of independent random variables and the central limit theorem is applicable to it. The distribution of μ is therefore approximately normal with parameters

$$E\mu = np,\ \sqrt{\mathrm{var}\,\mu} = \sqrt{npq}$$

which is indeed the De Moivre – Laplace theorem[7].

4.5. The application of the central limit theorem

Checking statistical homogeneity. In Chapter 1 we have discussed at length the statement that a scientific application of the theory of probability is conditioned by checking statistical homogeneity. Here, finally, we can explain the main pertinent methods.

The discussion usually concerns the following problem. In n_1 trials the event *A* occurred μ_1 times, in n_2 trials, μ_2 times. May we believe that the probability of success was the same in both series? Or, is the difference of the frequencies μ_1/n_1 and μ_2/n_2 sufficiently small and possible to be explained by purely random causes?

It is natural to assume that μ_1 and μ_2 are approximately normally distributed whether the trials were dependent or not. If the probabilities in the series are p_1 and p_2 then

$$E(\mu_1/n_1) = p_1,\ E(\mu_2/n_2) = p_2,$$

and, if $p_1 = p_2 = p$, $E(\mu_1/n_1 - \mu_2/n_2) = 0$. Also, it is natural to assume that the trials in both series are independent, then the magnitude ($\mu_1/n_1 - \mu_2/n_2$) should be approximately normally distributed with zero expectation and variance

$$\operatorname{var}\left(\frac{\mu_1}{n_1} - \frac{\mu_2}{n_2}\right) = \operatorname{var}\left(\frac{\mu_1}{n_1}\right) + \operatorname{var}\left(\frac{\mu_2}{n_2}\right).$$

If the terms in the right side are known, we could have said by means of a table of the normal distribution whether the mentioned difference can be explained by purely random causes or not. And for calculating those variances (but not at all for applying the central limit theorem) we have to assume that the trials in both series are independent, that is, that two series of Bernoulli trials were made whereas the central limit theorem does not demand complete independence. Then, if $p_1 = p_2 = p$ (whose value is unknown),

$$\operatorname{var}\left(\frac{\mu_1}{n_1}\right) + \operatorname{var}\left(\frac{\mu_2}{n_2}\right) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

It can be shown that the unknown p may be replaced here by

$$\hat{p} = \frac{\mu_1 + \mu_2}{n_1 + n_2}$$

and, assuming that the probabilities were identical, we see that

$$\xi = \left(\frac{\mu_1}{n_1} - \frac{\mu_2}{n_2}\right) \div \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

has an approximately standard normal distribution. Now, ξ can also be calculated only by issuing from numbers (μ_1, n_1) and (μ_2, n_2), i. e., by the known results of experimenting. It is known that the absolute values of ξ exceeding 2 or 3 are unlikely. It follows that, when obtaining values of that order, we should conclude that either the hypothesis $p_1 = p_2$ does not hold or, if it does, that an unlikely event had taken place.

How to choose between these conclusions? Some authors think that the decision theory can allegedly numerically express the risk of each possible choice and thus help here. However, the risk there is expressed by magnitudes which either are senseless or in any case will never be known to the researcher. The theory based on a quantitative expression of risk is always useless except in studies of games of chance[8]. Actually, the choice between the two mentioned decisions is a complicated procedure; at present, it is impossible to study it within the limits of a mathematical theory.

When solving such problems, we have to compare somehow the importance of each possible solution should it occur wrong. Both scientific and moral considerations denoted by the word *conscience* are involved here. Approximately similar but somewhat simpler is checking the hypothesis that the probability of success p in a given series of Bernoulli trials equals a given number p_0. If it is also true for a series of n trials with μ successes, then

$$(\frac{\mu}{n} - p_0) \div \sqrt{\frac{p_0(1-p_0)}{n}}$$

has an approximate standard normal distribution. Exactly by calculating that magnitude can the hypothesis of a male birth being equal to 1/2, see § 3.1, be checked (and rejected).

The arc sine transformation. I have just expounded the principles of checking the equality of probabilities in two series of Bernoulli trials. Now, I aim at indicating by an example the pertinent convenient methods developed in mathematical statistics. Nothing new in principle is here involved, but the practical convenience is essential. As an example, I choose the so-called *arc sine transformation* discovered by the celebrated English statistician Fisher. His idea was very simple: we consider some function $f(\mu/n)$ instead of μ/n, of the frequency of success itself. We have

$$f(\frac{\mu}{n}) = f[(\frac{\mu}{n} - p) + p] = f(p) + f'(p)(\frac{\mu}{n} - p) + \ldots$$

For large values of n, that frequency is close to p, so that we ignore the other terms of that expression and

$$\operatorname{var} f(\frac{\mu}{n}) = \operatorname{var}[f'(p)(\frac{\mu}{n} - p)] =$$

$$[f'(p)]^2 \operatorname{var}\frac{\mu}{n} = [f'(p)]^2 \frac{p(1-p)}{n}.$$

Let us choose the function f in such a manner that the expression for $\operatorname{var} f(\mu/n)$ would not depend on the unknown parameter p. More precisely, we assume that

$$[f'(p)]^2 p(1-p) = 1.$$

This is a differential equation and we may choose any of its solutions as $f(p)$; in particular,

$$f(p) = 2\arcsin\sqrt{p}.$$

Since, if allowing for the approximation made, $f(\mu/n)$ is a linear function of μ and, for large values of n, the distribution of μ is approximately normal, the expression

$$f(\frac{\mu}{n}) = 2\arcsin\sqrt{\frac{\mu}{n}} \qquad (4.8)$$

is also approximately normal. Its expectation is approximately

$$f(p) = 2\arcsin\sqrt{p}$$

and variance approximately $1/n$ which is how we have chosen the function f and it does not therefore depend on p. It also occurs that the distribution of (4.8) is even more close to the normal that that of the number μ itself.

Now let us have those two series of Bernoulli trials with n_1, μ_1, p_1 and n_2, μ_2, p_2 and suppose we wish to check the equality of the probabilities. Assuming that the two series are independent, the magnitude

$$2\arcsin\sqrt{\frac{\mu_1}{n_1}} - 2\arcsin\sqrt{\frac{\mu_2}{n_2}}$$

is approximately normal and its expectation

$$2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}$$

vanishes if the hypothesis is true. The variance of that random variable is the sum of the variances, $(1/n_1) + (1/n_2)$, and

$$[2\arcsin\sqrt{\frac{\mu_1}{n_1}} - 2\arcsin\sqrt{\frac{\mu_2}{n_2}}] \div \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (4.9)$$

has a standard normal distribution N(0, 1). After calculating (4.9) we may either adopt or reject the hypothesis of equal probabilities.

Mathematical statistics has plenty of such simple but very convenient methods; here, convenience is really attained when applying tables of the function $2\text{arc}\sin\sqrt{x}$ included in many collections of statistical tables; even a slide-rule will do.

Behaviour of the sum of independent random variables. When considering independent identically distributed random variables (4.4) with expectation and variance a and σ^2 respectively, their normed sum

$$s_n^* = \frac{\xi_1 + \ldots + \xi_n - na}{\sigma\sqrt{n}}$$

can not be especially large at any fixed n. Thus, in case of the normal distribution we easily find by means of its table that

$$P\{|s_n^*| \le 3\} = 0.997$$

and it is almost certain that

$$|\xi_1 + \ldots + \xi_n - na| \le 3\sigma\sqrt{n}. \qquad (4.10)$$

True, we ought to caution readers that that statement was derived for any fixed n. When considering a set $s_1^*, \ldots, s_n^*, \ldots$ we certainly can not state that all of its terms were less than 3 in absolute value. The

distribution of the maximal term $|s_k|$, $1 \le k \le n$, presents a special problem which we will not discuss.

For any fixed n the deviation of $\xi_1 + ... + \xi_n$ from its expectation na is only possible by a magnitude of the order of $\sqrt{n}$. This means that the non-random magnitude na, certainly if $a \neq 0$, plays a predominant part as compared with random deviations of the order $\sqrt{n}$. Now, dividing (4.10) by n, we get

$$|\frac{\xi_1 + ... + \xi_n}{n} - a| \le \frac{3\sigma}{\sqrt{n}} \qquad (4.11)$$

which is valid with probability 0.997 (if we believe in the normal distribution). When replacing 3σ by 4σ or 5σ, this inequality will be valid even with a higher probability. Given a large n, the difference in the left side of (4.11) is practically certainly small which is the celebrated law of large numbers.

It is interesting to dwell on the history of its proof and interpretation. It was note long ago that the results of separate observations, physical, meteorological, demographic or other, fluctuate essentially whereas the mean values of a large number of observations reveal a remarkable stability. The first statisticians had seen here divine intervention, but, as a scientific understanding of the world was being established, that stability became a scientific fact.

In the 18th, the *century of reason*, mathematics became very trusted; it was believed that the main laws of natural sciences and even of economics, moral philosophy and politics can be derived by that science. A desire to regard the stability of mean values as a mathematical theorem had been established and that opinion persisted in the 19th century. Exactly in that sense did Poisson interpret the discovery of his form of the law of large numbers and he thought that he had succeeded in proving that the mean of really made observations should be stable.

Chebyshev essentially developed the mathematical form of the law of large numbers by reducing its proof to the application of the [Bienaymé –] Chebyshev inequality. His proof can be found in any textbook on the theory of probability, and after him that law began to be considered as a very simple theorem independent of, and expounded before the central limit theorem. Students are now even taught to apply the inequality

$$P[|\frac{\xi_1 + ... + \xi_n}{n} - a| > \varepsilon] \le \frac{\sigma^2}{n\varepsilon^2}$$

for estimating the probability that the mean will deviate from a more than by ε. However, such an application (of the Bienaymé – Chebyshev inequality) is absolutely absurd because the central limit theorem provides a much more precise result. True, the Chebyshev form of the law of large numbers demands less mathematical restrictions to be imposed on the random variables ξ_i as compared with the central limit theorem, but it is just the same practically impossible

to check whether the appropriate mathematical restrictions are met. And it is certainly impossible to distinguish when only the Chebyshev theorem is valid from the case when both it and the central limit theorem are valid.

The evolution of the opinion on the natural scientific significance of the law of large numbers is connected with Mises. He especially indicated that there can not be any mathematic proof that the mean of the results of an experiment should be close to some number. Nowadays, we believe much less than at the time of Laplace and Poisson that the laws of the outer world can be mathematically derived. There exist too many causes which can change the course of an experiment from what should have followed according to our mathematical model.

For example, the conditions of all the known mathematical theorems on the law of large numbers include as an assumption that (4.4) is a sequence of random variables. Practically this means that we may discuss the probability of an event consisting in that ξ_1 took a value from some number set A_1, ξ_2, the same from A_2, etc, so that for any sets A_1, ..., A_n the event $\{\xi_1 \in A_1, ..., \xi_n \in A_n\}$ should be statistically stable. However, possible sets A_1, ..., A_n are so numerous that an experimental check of the stability of all such events is impossible. And the violation of statistical stability wholly depreciates any stochastic theorem and can be the cause of the observed violation of the stability of experimental means.

The natural scientific significance of the law of large numbers is now reduced to an understanding that when stochastic models are applied the corresponding theorems reflect the experimental fact of the stability of means. In Chapter 1 we indicated that there are many problems, for instance in geology or economics (their examples can be multiplied without any difficulty) in which it is senseless to discuss the statistical homogeneity of the ensemble of experiments. It is interesting that in such cases stability of means rather often also persists. We must acknowledge that we do not nowadays have any satisfactory mathematical explanation of the stability.

In the 20th century the study of the law of large numbers by means of a model of the space of elementary events had been essentially advanced. The so-called strong law of large numbers connected with Borel and especially Kolmogorov was discovered. For explaining its essence recall that in the Kolmogorov model the random variables (4.4) are functions $\xi_1(\omega)$, ..., $\xi_n(\omega)$ considered in the space of elementary events. It is possible to consider the event, that is, the set

$$\{\omega : \lim \frac{\xi_1(\omega) + ... + \xi_n(\omega)}{n} = a\}, \; n \to \infty$$

consisting of those elementary events ω for which that limit exists and is equal to a. The theorems of the type of the strong law of large numbers state that the probability of that set is 1 whereas the usual law does not deal with that set at all, it only discusses the sets of the type

$$\{\omega : | \frac{\xi_1(\omega) + ... + \xi_n(\omega)}{n} - a | \ > \ \varepsilon\}, \ \varepsilon > 0$$

for any finite n and states that the probability of such sets tend to vanish as $n \to \infty$.

After considering rather subtle mathematical examples it occurs that the strong law of large numbers is really strong: the ordinary law is certainly obeyed when the strong law is valid, but the inverse statement is not necessarily true. From the theorems concerning the strong law we indicate a very elegant Kolmogorov statement: for independent and identically distributed random variables the existence of expectation is sufficient for it to hold. Mathematically interesting is that the existence of variances is not demanded.

A special mathematical tool was needed for proving that theorem and in particular Kolmogorov discovered a remarkable inequality that goes under his name and generalizes the [Bienaymé −] Chebyshev inequality; the tool itself can certainly be applied in natural science. However, no such applications in which essentially more can be elicited from only the formulation of the strong law than from the usual law are discovered. This is connected with the fact that (see Chapter 2) a random variable $\xi = \xi(\omega)$ as a function of an elementary event is usually not observed; we know the value of $\xi(\omega)$, but not ω itself. We can rather discuss probabilities of various events. Similarly, it is somewhat senseless to discuss the observation of the limit of $\overline{\xi}$, we can only study $\overline{\xi}$ for a finite n. Those circumstances lead to any non-mathematical applications of the strong law being unlikely.

To conclude the problem of the application of the central limit theorem we will dwell on the statement made by no other but the undoubtedly great scientist of genius, Laplace, which for us is only interesting as being a psychological curious historical example. He discovered the mentioned above fact that for large values of n the sum $\xi_1 + ... + \xi_n$ behaves approximately like the non-random magnitude na whereas the random variations have order $\sqrt{n}$ so that with an increasing n that non-random magnitude will finally prevail over those variations. It follows that if $a > 0$, the sum of the random variables will also become positive.

Without any explanation Laplace infers that a colony situated far across the sea will finally achieve independence. He evidently imagined the strive for independence as some non-random factor whose action was gathering force with time whereas the opposite efforts of the metropolitan country as random variables with zero expectation. The first assumption is sufficiently understandable but the second one is very strange. However, in the long run Laplace was in the right: colonies did free themselves but we can not consider the effort to hold on to them as a random variable, it does not possess statistical stability. In the 19th century there was nothing special in sending out an expeditionary corps for putting down a rebellion in a colony but in our days that would have led to vigorous protests in the metropolitan country as well.

A scientist, discovering something remarkable (as the Laplacean central limit theorem) evidently can not keep from applying it everywhere. For example, in our time Wiener proposed to apply the theory of extrapolation of stochastic processes for forecasting the route of an airplane under anti-aircraft fire. That route however is not a stochastic process, or at least not such process for which there exists a theory of extrapolation and Wiener's proposal was senseless.

Evidently, science is collectively created; true, it is not beyond question whether an essential discovery can be made collectively, or is it necessary to have an outstanding scientist in a collective with its other members working in essence as his assistants. But what is undoubtedly a collective process is the delivery of science from the rubbish which some scientists usually adduce to their real discoveries.

4.6. When the central limit theorem can not be applied? That theorem is one of the reasons for believing that observational results usually obey the normal distribution. If only they, ξ_1, ..., ξ_n, are known, but not the parameters of the corresponding law, we are able to determine them approximately by appropriate methods. Indeed, according to the law of large numbers

$$a = \mathrm{E}\xi_i \approx (1/n)\,(\xi_1 + \ldots + \xi_n) = \overline{\xi}\,.$$

It can be shown that

$$\sigma^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2 = s^2\,.$$

The theory of errors allows to determine the precision of those approximate values.

In general, the observations are rather well describable by the normal law thus determined. In other words if

$$F(x) = P\{\,\xi_i < x\},$$

$\mathrm{N}(x,\ \overline{\xi}\,,\ s)$ being those probabilities calculated according to the normal distribution, then

$$F(x) \approx \mathrm{N}(x,\ \overline{\xi}\,,\ s).$$

However, this approximate equality is sometimes very perceptively violated. It happens when the values of x are such that $F(x)$ is near 0 or 1, – that its so-called tail areas are involved.

Let us begin by considering why those areas are practically significant in a special way. Suppose we intend to build some tall structure which will have to withstand high winds (or, if you wish, a spillway which has to pass spring floods, etc). We desire to reckon with such wind velocities that happen sufficiently rarely, once in a century, say. But how are we to find out that velocity? Or, if $\xi(t)$ is that velocity at moment t, we ought to indicate such a number x, that

$$P\{\max \xi(t) \geq x\} = 0.01,\ 0 \leq t \leq 1$$

where t is measured in years and the left part of the inequality is the maximal yearly wind velocity.

Suppose that we know the values $\xi_1, \xi_2, ..., \xi_n$ of the maximal velocity during the first, the second, ..., n-th year during which meteorological observations were made. However, wind velocities had not been recorded continuously but only several times a day, so that those maximal yearly velocities are in essence unknown. For the time being, let us nevertheless abstract ourselves from this extremely essential difficulty.

And so, we have those observations of the random variable ξ, the maximal yearly wind velocity, and we wish to assign an x such that

$$P\{\xi \geq x\} = 0.01. \qquad (4.11)$$

Had the number n been very large, we would have been obliged to select such an x that about a hundredth part of the ξ_i will be larger than it. The trouble, however, is that n, the number of years during which observations are available, is much less than 100. Then, if x is such that (4.11) is fulfilled, that is,

$$P\{\xi_i \geq x\} = 0.01 \text{ for each } i,$$

the number of variables ξ_i larger than x will obey the Poisson law with parameter $\lambda = 0.01n < 1$. It will follow that most likely all of our ξ_i will be less than x so that we are only able to say that x should be larger than each of the ξ_i's with no upper boundary available.

Therefore, we are tempted to smooth our $\xi_1, ..., \xi_n$ by some law, for example by the normal law $\mathrm{N}(x; \overline{\xi}, s)$ and determine x from equation

$$\mathrm{N}(x; \overline{\xi}, s) = 1 - 0.01 = 0.99.$$

Or, we will propose to identify the tail areas of the unknown function $F(x)$ with those of the normal law.

We turn the readers' attention to the fact that such a procedure should not be trusted either when applying the normal, or any other law, and that there exist both theoretical grounds and considerations based on statistical experiments for that inference. Theoretical grounds consist in that the central limit theorem only states that the difference between the exact distribution function $P\{s_n^* < x\}$ and the normal law is small:

$$P\{s_n^* < x\} - \mathrm{N}(x) \to 0.$$

For example, if that probability $P = 0.95$, $\mathrm{N}(x) = 0.99$ and the difference is only 0.04 which is sufficiently small. However, the relative error

$$[1 - P\{s_n^* < x\}] \div [1 - \mathrm{N}(x)] = 400\%$$

is very large. It is not indifferent that actually $P\{s_n^* \geq x\} = 0.05$ so that the event $\{s_n^* \geq x\}$ occurs once in 20 cases (once in 20 years, so to say) whereas by means of the normal distribution we found out that it happens once in a hundred cases (once in a century, so to say). We stress that the central limit theorem does not state that

$$[1 - P\{s_n^* < x\}] \div [1 - \mathrm{N}(x)] \to 1 \qquad (4.12)$$

uniformly for every x, and such a conclusion is actually wrong.

Thus, in the domain of probabilities close to 1 (and to 0) the application of the normal distribution can lead (and as a rule actually leads) to a large relative error whereas according to the central limit theorem the absolute error will be small. In particular, it should be borne in mind that the equality

$$P\{|\frac{\xi_1 + \ldots + \xi_n}{n} - a| \leq \frac{3\sigma}{\sqrt{n}}\} = 0.997$$

applied in § 4.5 is somewhat tentative. Instead of 0.997 values 0.990 or 0.980 can easily happen. Only when n is very large will 0.997 actually occur.

The ratio in the left side of (4.12) is stochastically studied by means of the so-called theorems of large deviations (Feller 1966). Their practical significance is however insufficiently clear. Incidentally, they indicate that the result will not be better if other frequently occurring distributions, for example, the Pearson curves, are applied instead of the normal law.

As to the available statistical experience of working with the tail areas of distributions, it shows that their behaviour is irregular. The violation of statistical homogeneity influencing the outcome of separate trials possibly especially concerns those areas. In such cases the attempts of describing the trials by statistical methods are hopeless.

The study of *the values of wind velocities possibly occurring once in a century* becomes complicated also because maximal yearly values are meant. If the values of those velocities at given moments are naturally assumed to be normally distributed, that maximal value will be naturally considered by means of some distribution of extreme values. However, these latter are only derived for independent magnitudes and are therefore unable to allow for a gradual increase of wind velocities under certain meteorological conditions. In addition, the theory of extreme values itself is often applied at an essential stretch. Recall also the lack of continuous records of wind velocities and you will be able to say absolutely for sure that nowadays there exist no scientific method of finding out how strong can the wind be *once in a century*. The designers should find some other method for stating how reliable are their buildings.

Notes

1. This example and considerations pertaining to medical statistics below are certainly in order. However, it is instructive that Soviet authors apparently avoided illustrations concerning touchy social statistics. O. S.

2. A drawing of lots described in the Talmud (Sheynin 1998) shows that the participants doubted the irrelevance of the order of drawing.

Laplace (1812/1886, p. 413) was apparently the first to note that preliminary drawings tend to equalize chances of the participants. O. S.

3. Sometimes experts have to apply subjective probability for various estimations. The same apparently may be said about jurors. Jakob Bernoulli, in his *Ars Conjectandi*, introduced non-additive subjective probabilities. He could have borrowed that idea from the scholastic theory of probabilism according to which the opinion of each Father of the Church was considered probable. O. S.

4. The author's conclusion is too harsh. Laplace and Poisson apparently only examined the ideal case; the former mentioned this restriction only in passing, the latter did not at all. Their work concerned general recommendations, for example, about the needed number of jurors. During the latest few decades the interest in stochastic studies of the administration of justice has been revived, although much more stress is now laid on interpreting background information (e. g., on estimating the number of possible perpetrators).

Laplace considered the juror's mistake (see somewhat below) according to the Bayesian approach and apparently only as a first approximation. See his actual understanding of that point elsewhere (Laplace 1812/1886, p. 523). Gelfand & Solomon (1973, p. 273) somewhat softened the issue of the interdependence of jurors. O. S.

5. The author cited the first Gauss' justification of the principle of least squares (which he later abandoned). Gauss arrived at the normal distribution by assuming, in part, that the arithmetic mean was the best estimator of a set of measurements. Incidentally, the *true value* mentioned by the author has been later understood as the limit of the appropriate arithmetic mean (Sheynin 2007). O. S.

6. The author could have stressed that a rigorous proof of the central limit theorem was only due to Liapunov and Markov, then to Chebyshev. I also note (Zolotarev 1999, p. 794) that that theorem is now understood in a somewhat more general sense (as the appearance of the normal distribution or its analogues). O. S.

7. It is in order to note additionally that De Moivre proved his theorem (the first proof of the most simple case of the central limit theorem) not at all as the author did. O. S.

8. The author did not, however, mention any such theory. O. S.

Bibliography

Arnold V. I. (1968), *Lektsii po Klassicheskoi Mekhanike* (Lectures on Classical Mechanics), pts 1 – 2. Moscow.

Bernstein S. N. (1926), Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Annalen*, Bd. 97, pp. 1 – 59.

Cournot A. A. (1843), *Exposition de la théorie des chances et des probabilités*. Paris, 1984.

Feller W. (1950, 1966), *Introduction to Probability Theory and Its Applications*. New York, vol. 1, 1950; vol. 2, 1966. Later editions available.

Feynman R. P., Leighton R. B., Sands M. (1963), *Lectures in Physics*, vol. 1, pt. 1. English – German edition. München – Wien – Reading (Mass.), 1974.

Gelfand A. E., Solomon H. (1973), A study of Poisson's models for jury verdicts in criminal and civil trials. *J. Amer. Stat. Assoc.*, vol. 68, pp. 271 – 278.

Gnedenko B. V. (1950, Russian), *Theory of Probability*. Moscow, 1969, 1973.

Gnedenko B. V., Kolmogorov A. N. (1949, Russian), *Grenzverteilungen von Summen unabhängiger Zufallgrößen*. Berlin, 1959.

Kolmogorov A. N. (1933, German), *Foundations of the Theory of Probability*. New York, 1950, 1956.

Laplace P. S. (1812), *Théorie analytique des probabilités. Oeuvr. Compl.*, t. 7. Paris, 1886.

Mises R. von (1928, German), *Probability, Statistics and Truth*. New York, 1981.

Poisson S.-D. (1837), *Recherches sur la probabilité des jugements en matière criminelle et en matière civile etc*. Paris, 2003.

Sheynin O. (1998), Statistical thinking in the Bible and the Talmud. *Annals of Sci.*, vol. 55, pp. 185 – 198.

--- (2007), The true value of a measured constant and the theory of errors. *Hist. Scientiarum*, vol. 17, pp. 38 – 48.

Zolotarev V. M. (1999), Central limit theorem. In Prokhorov Yu. V., Editor, *Veroiatnost i Matematicheskaia Statistika. Enziklopedia* (Probability and Math. Stat. Enc.). Moscow, pp. 794 – 796.

II

V. N. Tutubalin

Treatment of Observational Series

Statisticheskaia Obrabotka Riadov Nabliudenii. Moscow, 1973

Introduction

Facts are known to be the breath of the scholar's life. In our century of exact scientific methods, *observation* usually means *measure*, and facts which we have to deal with, are as a rule expressed in numbers. In any scientific establishment you will be shown long series of numbers also represented by graphs drawn by coloured pencils on squared paper. All of them are observational series. What benefit can 4we elicit of such coloured splendour whose collection demanded many long years of efforts by many authors?

Observational series often lead to some evident conclusion. Thus, after the introduction of antibiotics into medical practice, mortality from most infectious diseases sharply declined, but no mathematical treatment for such conclusions is necessary: the result speaks for itself. In other cases, however, conclusions can be not so unquestionable, and we have to apply statistical treatment and attempt to make them more reliable by mathematical methods.

It is important to imagine that in many cases the statistical treatment is beneficial but that perhaps even more often it is useless and sometimes even harmful since it prompts us to make wrong conclusions. Thus, antibiotics are useless in cases of viral infection. This booklet deals with instances in which statistical treatment is scientifically justified.

1. Two Main Mathematical Models of Observational Series

1.1. Why is statistical treatment needed? As stated in the Introduction, it can be not necessary at all. One more such example concerns the reliability of machinery. Suppose we discovered some preventive measure that obviously lowers the number of failures. Our observational series (for example, the number of failures over some years) certainly confirms the efficacy of our finding and we may be satisfied. Human nature, however, is incessantly wishing somewhat better; since there are less failures, we will wish to have none of them at all, so we propose another development and desire to confirm its efficacy by showing that the number of yearly failures will lower still more.

You can guarantee that this will not be so easy. When the number of yearly failures is small, it will be noticeably influenced by random causes. This does not yet mean that it can be studied by purely statistical methods, because their applicability demands the probably lacking statistical homogeneity [i]. However, such models allow to reach some important conclusions which we need to bear in mind. In addition, once a good technological result is already achieved, and we strive for a still better outcome, statistical homogeneity occurs rather often.

We will therefore assume that a stochastic model for the number of failures is valid and consider the check of efficacy of the innovation. When recognizing stochastic methods in general it is very natural to acknowledge the Poisson distribution of rare events as well, i. e., to apply the formula

$$P\{\mu = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

in which λ is the mean yearly number of failures. Suppose we have introduced the innovation at the beginning of a year and that during that year no failures have occurred whereas the mean number of them for the previous years was 2. May we conclude that the new development was effective?

That number, 2, was derived from previous statistical data and it does not necessarily coincide with the real value of λ, but for the time being we will disregard this circumstance. And so, $\lambda = 2$. Then the probability of a purely random lack of failures, or of $\mu = 0$, will be

$$P\{\mu = 0\} = e^{-2} \approx 1/7.$$

Therefore, if recognizing the innovation's efficacy, and awarding prizes to its inventors, the loss of money will have probability 1/7. Thus, 1/7 of all the employees proposing something useless, for example, perfuming the machinery, will get prizes. The trouble is not so much that the money will be lost, but rather that absolutely false viewpoints will be accepted. And so, perfuming of machinery is entered in search engines which do not distinguish between truth and rubbish and therefore find its way into general practice. Next year 1/7 of those who applied that *method* will once more be happy and publish pertinent rapturous papers with the unlucky 6/7 keeping silence because papers on setbacks can not be written[1]. That process intensifies as an avalanche; chairs of perfuming are established at universities, conferences organized, dissertations and textbooks compiled.

Such a picture although really sad is not a pointless abstraction since some pertinent examples are known, and we will provide some in the sequel. Considerations of that picture compels us, as we see it, to estimate in a new manner the merits of real science of that wonderful achievement, of collective intellect. *Sciences of perfuming* do emerge now and then, and even often, flourish (the more numerous are those participating the more reports about successes are made since 1/7 of them will become yearly successful) but do not live long.

Someone will always destroy them, and only the really valuable survives. The part played by stochastic methods in that self-purification of science is far from being the least important, although to declare that its role is exclusive will be nonsensical. We should be able to say whether the observed outcome can have been purely random[2].

However, just as any other science, mathematical statistics can have its own branches treating *perfuming*. We will consider the general

structure of statistical methods, discuss what is certain and what tentative there and on what premises are they founded.

1.2. The part played by mathematical models. Any statistical treatment must be preceded by a mathematical model of the phenomenon studied stating which magnitudes are random, which not; which are dependent, and which not, etc. Sometimes you will encounter a delusion that tells you that if any magnitude is not determinate (if its values can not be precisely predicted), it may be considered random. This is completely wrong because randomness demands statistical stability. Therefore, indeterminate behaviour is not generally speaking, randomness; or, if you wish, in addition to determinate and random there exist indeterminate magnitudes which we do not know how to deal with.

A mathematical model can include either determinate or random magnitudes, or both, but, as of today, not those last mentioned. The art of choosing a mathematical model therefore consists in approximately representing the indeterminate magnitudes appearing practically always as either determinate or random. It is also necessary that the values of the determinate magnitudes or the distributions of the probabilities of the random variables be derivable from the experimental material at hand (or available in principle).

Let us return to the determination of the efficacy of a new preventive measure. We have an observational series

$$\mu_1, \mu_2, ..., \mu_n, \mu \qquad (1.1)$$

where μ_i are the numbers of failures for the previous years and μ, the same for the year when the innovation is being tested. Where is the mathematical model here? In case of rare failures it is rather reasonable to assume that the series (1.1) is composed of random variables. However, when introducing that term, we oblige ourselves to state the statistical ensemble of experiments in which the variable is realized. Two paths are open: either we believe that the number of failures before the innovation was implemented are realizations of a random variable, or we imagine the results of many sets of machinery identical to our set working under the same conditions. In the first, but not necessarily in the second case the magnitudes

$$\mu_1, \mu_2, ..., \mu_n \qquad (1.2)$$

ought to be identically distributed. Or, assuming a Poisson distribution, we have in the first case

$$E\mu_1 = E\mu_2 = ... = E\mu_n = \lambda \qquad (1.3)$$

and, in the second case we may assume that

$$E\mu_1 = \lambda_1, E\mu_2 = \lambda_2, ..., E\mu_n = \lambda_n$$

where

$$\lambda_1, \lambda_2, ..., \lambda_n \qquad (1.4)$$

can differ.

Theoretically, the second case is more general and therefore, at a glance, more inviting, but we will see now that it does not lead to anything and should be left aside. Indeed, we have to know the value of the Poisson parameter $\lambda = E\mu$ for the number of yearly failures during the test of the innovation had it been ineffective. However, if there is no connection between the numbers (1.4), this parameter is not at all linked with our observations (1.2). And so, we are unable to determine λ. Then, when estimating (1.4) we should choose estimators $\hat{\lambda}_i$ based on a single realization (if we only observed one set of machinery) and we can only very roughly assume that

$$\hat{\lambda}_1 = \mu_1, ..., \hat{\lambda}_n = \mu_n.$$

Thus, when choosing a very general model, we are unable to determine its parameters which happens always. On the other hand, a particular model with equalities (1.3) is able to provide better approximation

$$\hat{\lambda} = \bar{\mu}. \qquad (1.5)$$

For the case of an ineffective innovation it is natural to assume approximately that

$$\lambda = E\mu \approx \hat{\lambda} = \bar{\mu}. \qquad (1.6)$$

This particular model enables us, in general, to solve our problem, but it has another disadvantage: it can be wrong. For example, aging of the machinery can lead to increase of the mean number of failures from year to year:

$$E\mu_1 < E\mu_2 < ... < E\mu_n < E\mu.$$

Here, equality (1.6) will underestimate the actual value of $E\mu$. Suppose that $\hat{\lambda} = 2$ but that actually $E\mu = 4$, then properly

$$P\{\mu = 0\} = e^{-4} \approx 1/55$$

instead of the result of our calculation, $P \approx 1/7$, see § 1.1, after which we will not admit that the innovation is effective although actually almost surely it is such.

We see that when constructing a statistical model we have to choose between Scylla and Charybdis, that is, between a general model, useless since we are unable to define its parameters and a particular model, possibly wrong and therefore leading to false conclusions. It is only unknown which is Scylla and which is Charybdis.

Suppose that we have adopted the particular model, i. e. declared that the magnitudes (1.2) are identically distributed random variables. Will this be the sole necessary assumption? No, since we badly need to know how large can be the error of the approximate equality (1.6). For example, if $\hat{\lambda}=2,$ can the real value of λ be 4? In other words, we should be able to calculate the variance of (1.5). It is equal to

$$\operatorname{var}\hat{\lambda}=\frac{1}{n^2}[\sum_{i=1}^{n}\operatorname{var}\mu_i+\sum_{i\neq j}\operatorname{cov}(\mu_i\mu_j)].$$

For the Poisson law

$$\operatorname{var}\mu_i = E\mu_i = \lambda$$

and, as a rough estimate, it is possible to assume $\operatorname{var}\mu_i = \hat{\lambda}=\overline{\mu},$ but we can not say anything abut the covariations. The available data are usually far from adequate for estimating it. Nothing is left than to suppose that the variables (1.2) are independent, i. e. to consider that the covariations vanish.

We thus arrive at a model of independent identically distributed random variables, that is, to a sample. The reader will probably agree that our considerations, if not logically prove that only a model of a sample is useful, are still sufficiently convincing in showing that it is difficult to tear away from the sphere of ideas leading to the model of a sample. It is therefore very popular and researchers are trying to work with it provided that its falsity is not proven.

The chapters of mathematical statistics devoted to samples are undoubtedly in its best and the most developed part. However, the model of sample is sufficiently (and even too) often wrong. We saw that if the machinery is aging, the observations (1.2) do not compose a sample. The same is true when a preliminary period is involved, when the work begins by eliminating defects after which the number of failures drops. Other causes violating the identity of the distribution of the variables (1.2) also exist.

Their independence can also be violated. For example, if a failure will lead to a capital repair with the replacement of many depreciated although still workable machine parts, a negative correlation between μ_i and the depreciated and worn-out μ_j will appear. If, however, the wear and tear of a machine part intensifies the depreciation of the other parts and no replacements are made, the appeared correlation will be positive.

When introducing models differing from a model of a sample, we should evidently specify their distinction by a small number of parameters determinable either theoretically or by available statistical data. Complicated models, as stated above, are absolutely useless. It is practically possible to allow for either deviations from the identity of distributions given by determinate functions or from independence provided that identity is preserved. We will now consider such models. It should be borne in mind that both these models and the model of sample are sufficiently tentative. True, if a model is proper, our

conclusions are derived in a purely mathematical way and therefore certain. However, on the whole everything depends on the model. Statistical methods are as certain (not more or less) as the conclusions of other sciences applying mathematical means, for example physics, astronomy or strength of material. In practical problems these sciences can provide guiding lines but can not guarantee that we have correctly applied them.

1.3. Model of trend with an error. In a mathematical model of an observational series something is always determinate and something random. We will consider a model in which that series

$$x_1, x_2, \ldots, x_n$$

is given by formula

$$x_i = f(t_i) + \delta_i. \qquad (1.7)$$

Here t_i is the value of some determinate variable specifying the i-th experiment, $f(t)$, some determinate function (the trend) and δ_i, a random variable usually called the error of that experiment. This situation means that *the Lord* determined the true dependence by $f(t)$ so that we should have observed $f(t_i)$ in experiment i, but that *the devil* inserted the error δ_i.

For example, $f(t)$ can represent one or another coordinate of an object in space as dependent on time, and x_i is our measurement of that coordinate at moment t_i. The devil's interference δ_i can certainly be determinate, random or generally of an indeterminate nature. Thus, the observed x_i can be corrupted by a systematic error so that $E\delta_i$ is not necessarily zero. We may assume that $E\delta_i = C$ and does not depend on i but it is also possible to consider $E\delta_i = \varphi(t_i)$ is a function of t_i. Still worse will happen if $E\delta_i$ depends on a variable u_i which we can not check. In neither of those cases statistical treatment can eliminate the errors.

However, a sufficiently thorough planning of the observations can allow us to hope that the errors will be purely random in the sense that statistical homogeneity is maintained and there is no systematic shift: $E\delta_i = 0$. More precisely, the systematic error will be sufficiently small and can be neglected. Such situations indeed comprise the scope of the statistical methods.

After recalling what was said in § 1.2 it becomes clear that most simple statistical assumptions should be imposed on the errors δ_i. Most often these errors are supposed to be independent and identically distributed. Normality is also usually assumed. Only one of their deviations from the model of sample was brought into use: it is sometimes thought that their variances are not equal to one another but proportional to numbers assigned according to some considerations. Or, it is assumed that such numbers w_i called weights of observations are known that

$$w_1 \operatorname{var} \delta_1 = w_2 \operatorname{var} \delta_2 = \ldots = w_n \operatorname{var} \delta_n = \sigma^2$$

and the variances are inversely proportional to the weights

$$\operatorname{var}\delta_i = \frac{\sigma^2}{w_i}.$$

I have described the assumptions imposed on the random component of our observations. Now I pass to their determinate component $f(t)$ otherwise called *trend*.

The most simple and classical case consists in that the function $f(t)$ is of a quite definite class but depends on some unknown parameters $c_1, c_2, ..., c_k$:

$$f(t) = F(t, c_1, c_2, ..., c_k) \tag{1.8}$$

where the function F is given by a known formula or an algorithm of calculation. For example, in case of the motion of an object in space those parameters can be understood as its coordinates and velocities at any definite moment; other, more opportune parameters can also be introduced.

Then any coordinate $f(t)$ will be uniquely determined by the parameters and the Newtonian laws of motion (if that object has no engine). The problem consists in determining estimates of the parameters $\hat{c}_i$ given observations (1.7). It is solved by the Gaussian method of least squares: the estimates are determined in such a way that the minimal value of the function

$$\sum_{i=1}^{n}[x_i - F(t_i; c_1, ..., c_k]^2$$

of c_i will be attained at point $(\hat{c}_1, ..., \hat{c}_k)$.

More often, however, is the case in which the real dependence $f(t)$ is unknown. Here also the equality (1.8) is applied but the function F is chosen more or less arbitrarily. Thus, a polynomial might be chosen and the method of least squares once more applied.

Such a non-classical situation when $f(t)$ is not known beforehand demands a more detailed analysis, see a concrete example in the next Chapter. Here, however, we describe an absolutely different model also applied for statistically treating observational series.

1.4. Model of a stochastic process. The main attention is turned to the isolation of the determinate component, the trend $f(t)$. The values themselves, x_i, of the observational series (1.8) are not random; random are only the additional magnitudes δ_i considered as errors, noise, and generally the devil's machinations. Another approach is possible with randomness being considered the main property of the series under study which we now denote by

$$\xi_1, \xi_2, ..., \xi_n. \tag{1.9}$$

Here, the most simple model consists in treating that set as a realization of an n-dimensional random variable. Such a model can be useful if the experiment providing it can be repeated many times over,

i. e., if many observational series can be obtained under similar statistically homogeneous conditions. More often, however, we have only one such series, distributions of probabilities certainly can not be reconstructed and the model of an *n*-dimensional distribution is absolutely useless. However, if we assume that the joint distribution of the magnitude ξ_1, ξ_2, is the same as that of ξ_2, ξ_3, of ξ_3, ξ_4, etc, then the pairs (ξ_1, ξ_2), (ξ_2, ξ_3),..., (ξ_{n-1}, ξ_n) provide many realizations, although perhaps not mutually independent, of that bivariate distribution. Such a distribution is therefore determinable in principle.

It is convenient to generalize somewhat the mathematical model. Let us consider a sequence of random variables infinite in both directions

$$... \xi_{-1}, \xi_0, \xi_1, \xi_2, ..., \xi_n, \xi_{n+1}, ... \tag{1.10}$$

called a *stochastic process*. We assume that theoretically there exist distributions of probabilities of any finite set

$$\{\xi_\alpha, \xi_\beta, \xi_\gamma\} \tag{1.11}$$

of random variables. Our observational series (1.9) is a part of the infinite sequence (1.10) and only allows us to reach some conclusions about that whole process if the model includes a rule representing distributions of magnitudes (1.11) with negative and large positive subscripts through the distribution of the observed variables (1.9). Without such a rule the model of a stochastic process is useless.

In the most simple and most natural case the condition of stationarity is imposed: for any τ the distribution of the variables $(\xi_{\alpha+\tau}, ..., \xi_{\gamma+\tau})$ coincides with that for $\tau = 0$. The model of a stochastic process consists in that [now] we consider our observations (1.9) as a part of the realization (1.10) of a stationary stochastic process.

When assuming a model of a stochastic process, only bivariate distributions are usually applied and in addition only the correlation between the different values of that process are studied. It ought to be said that in spite of the popularity of the concept of stochastic process, only quite a few examples can be cited in which it allowed to describe adequately the statistical properties of observational series. Most publications begin by stating that a pertinent *stochastic process specified in such and such a way is given*, but there really are only a few works where these specifications are indeed determined theoretically or experimentally.

The theory of stochastic processes is here suitable for solving abstract problems: what will happen if a white noise of a given intensity influences some system. Such problems, however, only indirectly bear on the real behaviour of a system because under real conditions it is not likely the white noise that influences the system, – it does not even concern a stochastic process (lack of statistical homogeneity). But meanwhile often no one studies what is really acting on the system because such investigations are complicated, difficult and expensive so that it is much easier to restrict the attention to arbitrary prior assumptions.

It is interesting therefore to see what occurred when the most eminent statisticians attempted to study actual data by models of a stochastic process. Rather often they experienced failure, see Chapter 3. We will also briefly mention the statistical theory of turbulence in which the notion of stochastic process has been applied with brilliant success.

2. The Method of Least Squares

Gauss discovered and introduced it into general usage. The classical case which he considered consisted in that some known relations should be maintained between the terms of the observational series

$$x_1, x_2, \ldots, x_n$$

had not the observations been corrupted by errors. For example, in the case of the path of an object in space[3] it would have been possible to express all terms of the series through a few of its first terms had these been known absolutely precisely. This classical case can be comparatively easily studied within the boundaries of mathematical statistics. Practical applications of the method of least squares can encounter more or less essential calculational difficulties which we leave aside. Other difficulties are connected with the possible non-fulfilment of the assumption of the model of trend with error. Thus, errors of successive measurements of distances by radar apparently can not be assumed independent random variables. It is in general unclear whether they possess a statistical character so that statistical methods are here unreliable and moreover helpless.

The observations themselves, however, are highly precise and can be made many times, so that statistical methods are not needed there. In spite of all the merits of the classical case, its shortcoming is that it occurs comparatively rarely. Much more often we are convinced that our observations can be approximated by a smooth dependence

$$x_i \approx f(t_i)$$

where t_i is a variable describing the conditions of the i-th experiment. The exact form of the function $f(t)$ is, however, unknown.

Methods strongly resembling those of the classical case are applied here, but their study indicates that they are not mathematically justified. Mathematical statistics widely applies mathematics but is not reduced to that comparatively very transparent science. Statistics is rather an art and as such it has its own secrets and we will indeed begin by studying them.

2.1. The secrets of the statistical art. When wishing to apply the method of least squares we can in most cases use a computer programme compiled once and for all. It is just necessary to enter the data, wait for the calculations to be made and the printer will provide a formula for a curve fitting the observations. However, he who passes all these procedures to a machine will be wrong. It is absolutely necessary to represent the available data in a visible way and at least to glance at the figure.

The human eye is able to detect such special features in the material that the machine will miss. For example, if the first half of the observations is situated above, and the second half, below the fitting curve, then, obviously, the assumption of independent errors in the model of trend with error is violated. In such a case no computer calculation has any sense [since] the machine is unable to note these special features all by itself. A pertinent programme can certainly be compiled but the trouble is that there are so many possible features of the data for including the study of all of them in the programme.

It is natural to entrust the application of any given statistical test to a machine, which however is barely able to formulate the necessary tests. This should be done by a statistician by issuing from a visual estimation of the statistical data that should be therefore represented in a graphical way. It follows therefore that the statistical art is based in the first instance on visual estimation.

He who wholly trusts the automatic computer calculations deprives himself of the possibility of checking the statistical model and, as a result, the more is given over to machine treatment, the less trust it deserves. However, if statistical material demands to be estimated by the naked eye, this will be possible for functions of one variable well enough (since they can be represented by graphs), much worse with functions of two variables (they can be depicted by isolines like the heights above sea level on topographic maps) but we are absolutely unable to study functions of a larger number of variables.

That is the domain where we may only reckon on help from the computer. First steps were done here. Such directions like multivariate statistical analysis and design of extremal experiments have emerged, but it is still a very long way to go before really reliable methods are created. The methods of the directions just mentioned are sometimes effective, sometimes not and we do not know the reason why. The main shortcoming here is the low moral level of research, the custom of pretending the desired to be real so that we do not know what exactly can we trust in.

And so, when desiring to apply the method of least squares, we should begin by drawing a graph of the observational series. The reader will imagine what transpires here by having a look on the broken line on Fig. 1; its meaning is yet unimportant. *Such a broken line obviously fluctuates about some smoothly changing curve. This curve is indeed expressing the true regularity whereas the fluctuations of that broken line are occasioned by random causes and have no relation* [...]

The italicized phrases usually comprise all the available information about the real studied dependence. Understandably, it is too diffuse and indeterminate for directly admitting some scientific investigation. Conclusions reached by a naked eye study should be transferred into a mathematical model applicable for statistical treatment. That transformation is the second mystery of the statistical art.

In case of the method of least squares most often a model

$$x_i = P(t_i) + \delta_i,\ i = 1, 2, ..., n$$

is applied with $P(t)$ being a polynomial whose coefficients should be estimated by that method. [...] It is usually said that in case another model

$$x_i = f(t_i) + \delta_i,\ i = 1, 2, ..., n$$

is valid with $f(t)$ not being a polynomial, we may apply the Weierstrass theorem according to which we can approximate $f(t)$ by the polynomial $P(t)$ as precisely as desired. However, that reference is , however, inappropriate because any continuous function can be approximated by a polynomial of a sufficiently high degree. In practice we attempt to choose a low rather than a high degree. [...] A polynomial of a higher degree [can be] further from reality than that of a lower degree. Then, the Weierstrass theorem is also valid for functions of several variables. However, if the earth's surface as shown by isolines on a topographic map is considered a typical function of two variables, and its approximation is attempted, the result will be usually unsatisfactory: the degree of the polynomial should be too high. The theoretical Weierstrass theorem and practical smoothing differ.

The method of smoothing by a polynomial is therefore not mathematically justified and the success of that procedure is one more mystery of the statistical art. How can we explain the rather often success here? Apparently the human eye feels well enough the behaviour of the graphs of analytic functions, polynomials in particular. As students are taught, only a few points ought to be calculated, – discontinuities, extrema, sometimes points of inflexion, – and functions are then reconstructed quite well. It may be supposed that we are able to catch whether the real dependence is approximated when smoothing a broken line by a polynomial well enough. This statistical procedure of smoothing a function by a polynomial is probably only applied when success is expected after having a look at the graph by naked eye.

The situation changes at once if that procedure is attempted to be wholly accomplished automatically. In that case no data will be preliminarily estimated and the portion of successful smoothing will be sharply reduced. The case of functions of many variables is quite complicated. We are unable to show either the experimental data or the result of smoothing and can not even say whether it was successful or not.

2.2. Smoothing by a polynomial: an example. It is time to explain the provenance of the observations represented on Fig. 1. [It was the study of the damage of insulation of the stators of large turbo-generators (Belova et al 1965, 1967).] The total number of failures is naturally comprised of failures of separate generators. Rather early in our work we decided that the probability of a failure of a given generator is proportional to the total area of its insulation and little depends on its constructive or operational conditions; hydrogen cooling was then almost non-existing. We have therefore studied the behaviour of a unit area of that insulation (100sq. m. corresponding in its order to an area of insulation a large machine) without allowing for any other peculiarities. The problem of aging of the insulation, i. e., of

the increase of the probability of failure with time, was formulated. [...]

The values of the frequencies of failures comprise a broken line. Their scatter increases with t, a circumstance connected with a sharp decrease of the area of insulation, i. e., of the amount of experimental material.

We are interested in the values of probabilities $p(t)$ of a failure of a unit area of insulation aged t during unit time (10^4 working hours, about 1.5 years). For small values of t the amount of experimental material is large, but $p(t)$ themselves are low, 0.01 – 0.02, so that their direct determination through frequencies is fraught with very large errors. The mean square deviation of the frequency, μ_i/S_i, where μ_i is the number of failures during time interval between $(i-1)$-th and i-th time units and S_i, the corresponding area of insulation, is known to be equal to

$$\sqrt{\frac{p(t_i)[1-p(t_i)]}{S_i}}$$

where $t_i = 10^4 i$ hours. For $t_1 = 10^5$ $p(t_i) \approx 0.02$, $S_i = 200$, so that deviation is roughly 0.01 or 50% of $p(t)$ itself.

Then it is natural to attempt to heighten the precision of determining $p(t)$ by smoothing since the estimation of this probability then depends on all other experimental data. But then, a statistical model is necessary here. It is rather natural to consider the observed number of failures (1.2) as random variables with a Poisson distribution. Understandably,

$$\mathrm{E}\mu_i = S_i\, p(t_i).$$

It is somewhat more difficult to agree that the magnitudes μ_i are independent. Here, however, the following considerations applicable to any rare events will help. Take for example μ_1, μ_2. Failure occurring during the first interval of time influences the behaviour of the insulation in the second interval, but that action is only restricted to the failed machines whose portion was small. Having admitted independence, the mathematical model is completely given although it is connected not with the most convenient normal, but with the Poisson distribution. Then, the variances

$$\mathrm{var}\ \mu_i = \mathrm{E}\mu_i = S_i\, p(t_i)$$

depend on probabilities $p(t_i)$ which we indeed aim to derive. A transformation to magnitudes

$$v_i = 2\sqrt{\mu_i} \qquad (2.1)$$

essentially equalizes the variances and therefore helps.

These magnitudes $v_1, v_2, \ldots, v_n$ from which we later return to magnitudes (1.2) are smoothed. The smoothing itself is easy in essence

but some mathematical tricks described in detail elsewhere (Belova et al 1965, 1967) are applied.

The approximate expression

$$p(t_i) = p(x_i) \approx (1/4)[b_0 - 0.1333b_2 + b_2x^2]^2 + 0.35/S_i,\ x_i = t_i/22,$$

where t_i is measured in the selected intervals of time and the last term is necessary for allowing for the systematic error that occurred when transferring to v_i (2.1) should be considered final.

Estimates for b_0 and b_2 and their variances are

$$\hat{b}_0 = 0.225,\ \hat{b}_2 = 0.20,\ \ \mathrm{var}\,\hat{b}_0 = 2.12\cdot10^{-4},\ \ \mathrm{var}\,\hat{b}_2 = 44.5\cdot10^{-4}.$$

The magnitudes $0.35/S_i$ are smoothed by a certain polynomial.

Careful statistical work concerning probabilities of failures should apply our answer exactly in the provided form. However, it is not vivid enough and we have therefore represented it in a simplified way. A confidence rectangle for (b_0, b_2) with an 80% coefficient was therefore indicated with curves $p_1(t)$, $p_2(t)$, $p_3(t)$ added. Curve $p_2(t)$ provides the best estimate of the real probability of failure which we are able to offer. It corresponds to the estimates of b_0 and b_2. The other curves are obtained if the real point (b_0, b_2) is replaced by the left lower and right upper vertices of the confidence rectangle respectively. They provide an idea about the order of the possible error of curve $p_2(t)$ but, strictly speaking, are not the boundaries of the confidence region for the true curve. The confidence region for $p(t)$ can be constructed in different ways. Strictly speaking, it is not needed since all the information applied for constructing it is summed in the mentioned variances, $\mathrm{var}\,\hat{b}_0$ and $\mathrm{var}\,\hat{b}_2$. The region between $p_1(t)$ and $p_3(t)$ can be considered as some approximate (having the adequate order) version of the confidence region.

Various versions of checking our model by statistical tests are of fundamental significance. In statistics, no check is exhausting but a number of well passed tests nevertheless produces a feeling of certitude in the results. We will dwell in detail on all these checks.

The simplest criterion is the study of the final result. Let us have a good look at the curve representing the dependence sought. Do not the actual data deviate too much? (The maximal among the 22 deviations is $\mu_{20} = 4$ and according to the Poisson formula $P\{\mu_{20} \geq 4\} \approx 0.12$.) In itself, this is not especially significant, and for the maximal of 22 deviations with only $1/22 \approx 0.05$, it is quite acceptable. [...]

It is possible to compile an expression similar to the sum of the squares of deviations of the experimental data from the smooth curve $p_2(t)$. But it is better to deal with magnitudes v_i (2.1) since their variance does not essentially depend on the unknown probabilities $p(t_i)$ and is roughly equal to 1. In our case, the variance of the observations is thus known almost exactly, a circumstance connected with the Poisson distribution, which depends only on one parameter rather than two as the normal law does. In general, the test applying the sum of the squares of deviations also shows that the final curve fits well enough.

Another group of tests is connected with the choice of the degree and the number of terms of the approximating polynomial. Here, we also deal with v_i (2.1) and test what happens when they are approximated by various polynomials up to the third degree inclusive. It is obvious that the polynomial sought includes a free term. Then we add, in turn, terms of the first, second and third degree. The best improvement of approximation is reached when polynomials of the type

$$c_0 + c_2 t^2 \qquad (2.2)$$

are chosen.

And now we check that the addition of terms of the first and third degree to it does not significantly improve the approximation; for details, see Belova et al (1965, 1967). After all these checks we become sure that applying a polynomial (2.2) we have indeed as completely as was possible elicited the determinate component from the available data.

However, having happily concluded the tests of the hypotheses connected with the smoothing, we do not at all check the main hypothesis, that the probability of the failure of a unit area of insulation does not depend on the constructive or operational peculiarities of the pertinent machine. Indeed, we are only checking whether the magnitudes μ_i are obeying the Poisson distribution (and, in part, whether they are independent).

However, that distribution also occurs when the probabilities of failures occurring on different areas of the insulation are unequal (provided all the probabilities are sufficiently low). The most important hypothesis of statistical homogeneity of the various unit areas of insulation is yet left unchecked and can not be checked by issuing from the generalized data of Fig. 1[4]. At the same time most interesting is exactly the isolation and study of machines with high and low break-down rates (or a confirmation that all of them have the same rate of failures). We will see now how these problems can be solved.

2.3. Check of statistical homogeneity. The most important condition of acquiring a statistically homogeneous totality, or, so to say, the most important mystery of the statistical art consists in carefully selecting the material to be studied. Thus, the data of Fig. 1 does not include failures of the insulation occurring because of causes [of various causes of its random damage]. We supposed that such causes, although usually called random, are not random in the stochastic sense since they are not statistically stable.

The selection of material was made easier by the fact that a failure of a large machine is an extreme event whose causes are thoroughly investigated and duly registered. The most suitable for including a failure into statistical treatment was the formulation *local defect of insulation*. In general, however, all failures were included if an alien cause was not clearly indicated. Failures included into statistical treatment composed about a half of all the failures of insulation. When selecting material, the statistician must invariably keep to some principle once and for all.

It is clear therefore that no special significance can be attached to statistical calculations of reliability. This conclusion is important for a principled evaluation of the real meaning of the reliability theory. Now, however, our interest is concentrated on another point, on ascertaining whether our thoroughly selected totality was statistically homogeneous. Suppose that practically the derived curve precisely expresses the probability of failure, $p(t)$. If the failures of the insulation are mostly due to its local damage, it is logical to assume that a failure of a certain machine does not influence (or little influences) its failure after repair.

But then the total number of failures ξ_i during all the operational time of a machine is a sum of independent random variables, – the number of failures during the first, the second, ... selected intervals of time. Each term obeys the Poisson distribution, so that the total number of failures also obeys it. The parameter of that distribution for machine i for the $(k-1)$-th time interval is

$$\lambda_{ik} = p(t_k)S_i \approx p_2(t_k)S_i \qquad (2.3)$$

where as before S_i is the area of insulation of machine i.

Therefore, the parameter

$$\lambda_i = \mathrm{E}\xi_i \qquad (2.4)$$

of the total number of failures for the i-th machine can be calculated by summing the expressions (2.3) over such t_k that are less than the general working time of the pertinent machine. We may thus consider that the numbers (2.4) are known for all the machines. [...] This method of determining λ_i is only valid when statistical homogeneity is supposed, otherwise the computed curve $p_2(t)$ only provides a general characteristic of the breakdown rate.

Some machines will have a higher, other machines, a lower rate, – will have either more or less failures than indicated by the Poisson law with parameter calculated according to our rule. So it seems that we have established the effect to be sought for in order to check violations of statistical homogeneity. However, the trouble is that it is very difficult to discern that effect. Indeed, suppose we have determined that for a certain machine $\lambda_i = 0.1$ whereas $\xi_i = 2$. Since

$$P\{\xi_i \geq 2\} = \frac{\lambda_i^2}{2} e^{-\lambda_i} + \ldots \approx \frac{1}{200}$$

it would seem that we detected a significant departure from that homogeneity. But statistics covers several hundred machines, so that for one (and even for a few) of them an event with probability 1/200 can well happen.

There are several possible ways for establishing a useful statistical test of homogeneity. One of them is, to apply the Poisson theorem once more. Consider the total number of machines that experienced one, two, three, ... failures. We will show that the distribution of

probabilities for those magnitudes can be derived. Introduce a random variable

$f_k(\xi) = 1$, if $\xi_i = k$; 0, if not, $k = 1, 2, 3, ...$

Since ξ_i is the number of failures for the i-th machine, the number of machines that had k failures is equal to $\sum f_k(\xi_i)$, a sum of independent random variables. For most machines λ_i is near zero, therefore, if $k \neq 0$, the probability

$P\{f_k(\xi_i) = 1\}$

is low, and the sum above roughly obeys the Poisson distribution. Its parameter is derived from

$$\mathrm{E}[\sum_i f_k(\xi_i)] = \sum_i \mathrm{E} f_k(\xi_i),\ \mathrm{E} f_k(\xi_i) = P\{\xi_i = k\} = \frac{\lambda_i^k}{k!} e^{-\lambda_i}$$

where, provided the hypothesis of statistical homogeneity is valid, λ_i is calculated as stated above. A simple calculation (Belova et al 1965, 1967) indicates that at different values of k the studied sums are close to independent random variables.

How does deviation from statistical homogeneity reveal itself? Some machines will have a higher breakdown rate and experience two or more failures, other will deviate in the opposite sense and work failure-freely. When statistical homogeneity is corrupted, the number of machines with two or more failures will increase, and will decrease for those with one failure.

The treatment of actual data resulted in the following number of machines with 1, 2, 3 and 4 failures (line 1) as compared with the corresponding expectations (line 2).

1.	27	10	1	1
2.	29.6	5.7	1.5	0.44

The number of machines with one failure decreased insignificantly but of those with two failures increased noticeably: for the Poisson law with parameter 5.7 the probability of 10 or more is 0.065. For $k = 3$ and 4 the deviations were small.

The only deviation worth discussing is that for machines having 2 failures. However, we may consider it maximal for four independent deviations and then its probability is $1 - (1 - 0.065)^4 \approx 0.25$ so that its deviation is not especially significant.

Although the hypothesis of statistical homogeneity had passed a rather rigid test with credit, some shadow of doubt is still cast on it. This seems to mean that for most machines the breakdown rate is roughly the same but that small groups of them it can stand out. A wide scatter would have led to an essentially more significant result of the test. [...]

It follows that in general the derived fitting mean curve $p_2(t)$ can be applied for an approximate calculation of the mean number of failures

of various groups of machines, and this provides us a test for estimating the reliability of the insulation. Purely statistical methods certainly do not concern the improvement of that reliability which is a technological problem. But at least we may say whether the reliability of insulation had changed and in which direction or that it remained as it was previously. This is the practical significance of the work done which would not be so important had the comparatively high statistical homogeneity of the insulation not been established. [...]

2.4. The naked eye study. We had assumed that smoothing by polynomials is usually successful because the data for that treatment is selected beforehand by naked eye. It would have been improper to fail to mention that physicists and engineers also often perform the smoothing itself by naked eye without applying the method of least squares. And how do we decide that the smoothing in a given case was successful? Perhaps because the curve derived by least squares passes exactly where it would have been if drawn without applying that method?

An experimental smoothing by naked eye of the broken line in § 2.2 was carried out. Participants were mathematicians, workers at a statistical laboratory, and engineers. Each received a list of paper with only that line shown [...]. The results achieved by an overwhelming majority were very good. Fifteen out of sixteen of those participating had almost completely drawn their curves between the two curves, $p_1(t)$ and $p_3(t)$ as shown on Fig. 2. [...]

I. V. Girsanov, the chief of one of the sections of the statistical laboratory, achieved the best result; he unfortunately perished in a later tourist mountain tour. [...] In general, the results of smoothing by naked eye are quite comparable in precision with the method of least squares. Had we been only interested in curve $p_2(t)$, we could have well drawn it without any calculations. However, a thorough statistical treatment demands an estimation of precision as well for which a statistical model and science in general are necessary.

Thus, when estimating the probability of success in Bernoulli trials, we turn to frequencies, but for understanding how large can the deviations of frequency from probability be, we should, first, consider the trials independent (the statistical model) and second, apply the De Moivre – Laplace theorem which (however done) is proven in a complicated manner [in essence, by the former in 1733] and this is undoubtedly science.

When smoothing a broken line by naked eye, we do not even have to know the number of observations used for calculating its points [...] and anyway it is impossible to indicate the confidence region for the curve sought. Here, we need all the science connected with the method of least squares and still the almost complete coincidence of the area shown on Fig.2 with that between the curves $p_1(t)$ and $p_3(t)$ demands to be somehow explained.

Note, however, that for small values of t that first area is somewhat narrower than the second one whereas that latter, as shown by calculation, is 1.5 – 2 times narrower there than a thoroughly constructed confidence region with the usual confidence coefficient of 0.70 – 0. 95. This means that the indefiniteness of the naked eye

smoothing is, however, in general less than it is when calculated according to the rules of statistics.

It occurs because, when deciding by naked eye, we have to do with a given graph, with a result determined by random experimenting; on the other hand, when working by statistical methods, we apply a statistical model and therefore also cover the possible scatter of the results of random experiments themselves from one of their realizations to another one. However, the general problem of the real possibilities of the naked eye methods demands wide experimental investigation.

3. The Theory of Stochastic Processes

However beneficial (in suitable cases) is the method of least squares, a glance at the observational series often convinces us that the model of trend with error can not describe the observations, since we are unable to isolate by naked eye a determinate curve with observational points chaotically scattered around it. This is what Slutsky (1927/1937, p. 105), a co-creator of the theory of stochastic processes, wrote about it:

Almost all of the phenomena of economic life, like many other processes, social, meteorological, and others, occur in sequences of rising and falling movements, like waves. Just as waves following each other on the sea do not repeat each other perfectly, so economic cycles never repeat earlier ones exactly either in duration or in amplitude. Nevertheless, in both cases, it is almost always possible to detect, even in the multitude of individual peculiarities of the phenomena, marks of certain approximate uniformities and regularities. The eye of the observer instinctively discovers on waves of a certain order other smaller waves, so that the idea of harmonic analysis [...] *presents itself to the mind almost spontaneously.*

The *idea of harmonic analysis* can nevertheless attempted to be achieved by the model of trend with error. It is done by the so-called method of *periodogram* that preceded the methods of the theory of stochastic processes and we will briefly consider it.

3.1. The periodogram method. Suppose that our observations made at discrete moments of time, each second, say, can be described by the model

$$x_t = \sin(\lambda_0 t + \varphi) + \delta_t,\ t = 0, 1, ..., n \qquad (3.1)$$

where λ_0 is some parameter (circular frequency of oscillation), φ, the phase of oscillation and δ_t, random error. Suppose that λ_0 is much less than 2π, so that successive observations of only one component $\sin(\lambda_0 t + \varphi)$ would provide a clearly seen sine curve each unit of time (which is much shorter than the period of oscillation, $2\pi/\lambda_0$). The addition of random errors (suppose, for the sake of simplicity, independent) will certainly corrupt the picture. So how to reconstruct the frequency λ_0?

Multiply our observations x_i by $\sin(\lambda t)$ and $\cos(\lambda t)$ where λ is a variable, and consider the sums

$$A(\lambda) = \sum_{t=1}^{n} x_t \sin \lambda t,\ B(\lambda) = \sum_{t=1}^{n} x_t \cos \lambda t.$$

In previous times this calculation for various values of fairly many λ was rather tedious, but computers removed that difficulty. Calculate now the function

$$C(\lambda) = A^2(\lambda) + B^2(\lambda)$$

called periodogram. Formerly, it was imagined as a function of the period, $2\pi/\lambda$, rather than of frequency λ, which explains the origin of that term.

The main statement is that, given a sufficiently large number of observations n, the periodogram as a function of λ will take a clearly expressed maximal value in a small vicinity of the real frequency λ_0. If the determinate part of the observations consists of several harmonics $\sin(\lambda_0 t + \varphi)$ rather than one; that is, if

$$x_t = \sum_{j=0}^{k} A_j \sin(\lambda_j t + \varphi_j) + \delta_t, \qquad (3.2)$$

then the periodogram will have several maximal values situated close to $\lambda_0, ..., \lambda_k$. Their heights will depend on the number of observations, n, and amplitudes, A_j. When not knowing beforehand the number of harmonics and the variances σ^2 of the errors δ_i, we find ourselves in a rather difficult situation. The periodogram generally has very many local maxima and it is incomprehensible how to interpret them, either as really corresponding to *latent* periods λ_j or as occurring purely randomly[5].

These difficulties can be somehow overcome. It is worse that as a rule there is no guarantee that model (3.2) is valid. We can likely be sure that it is not. For example, when studying series of observations taken from economics, it is seen by naked eye that they rather smoothly depend on time (they rarely change from increasing to decreasing or vice versa) which should not occur for observations represented by model (3.2): they ought to be scattered around the smooth curve, the main term at the right side of (3.2). We may certainly assume that that curve itself badly corresponds to our idea of a smooth curve and that its roughness compensated random scatter, but an unreasonably large number of harmonics is needed for that to happen.

The most important problem therefore consists in determining how reasonable are the results provided by the periodogram method when the model (3.2) is wrong and the observational series $\{x\}$ is described by some other model. The generally known English statistician M. G. Kendall [Sir Maurice Kendall] carried out such experimental studies described in his rather rare book (1946) on mathematical statistics. This small contribution is one of the most remarkable books on mathematical statistics. Its epigraph is curious:

To George Udny Yule

To borrow a striking illustration from Abraham Tucker, the substructure of our convictions is not so much to be compared to the solid foundations of an ordinary building, as to the piles of the houses of Rotterdam which rest somehow in a deep bed of soft mud.

J. A. Venn, *The Logic of Chance* [1886][6]

We (§ 1.1) stated that *sciences of perfuming do emerge* [...] *flourish* [...] *but do not live long*. This had indeed happened to the periodogram method which was ruined in particular by Kendall (1946). He considers the model of autoregression (we will soon deal with it) which is as applicable as model (3.2) if not to a greater extent to analyzing series in economics. And in case of that model the periodogram method isolates frequencies that have absolutely nothing in common with its structure. Kendall concludes his opinion about that method in a brief sentence: *As misleading as it could be*.

It seems in particular that exactly in the same way Kendall regards the works of the renown English economist Beveridge who is celebrated due to his compilation and analysis by the periodogram method a few long series in economics, for example of cost of wheat in Europe covering 370 years. It could have been interesting to know the considerations that had guided him while compiling that series and whether it was done properly, but this is likely impossible. Beveridge compiled a periodogram and isolated many periods in his series which are likely senseless.

3.2. Stochastic processes. Nowadays correlation and spectral theories of stationary stochastic processes are applied instead of periodograms. A stochastic process is a function of variable t often but not necessarily playing the part of time and of an elementary random event ω. We will denote a stochastic process in an abbreviated form as ξ_i leaving aside the random argument ω since the functional dependence of the stochastic process on w is never considered in applications. Had we desired to describe clearly the space of elementary events, $\Omega = \{\omega\}$, the separate elementary events would have been as a rule extremely complicated. Thus, a separate elementary event is often understood as a function $\omega = \omega(t)$ of argument t.

In that case, the value of the stochastic process at moment t and elementary event ω is $\omega(t)$ which is a tautology pure and simple and practically does not lead anywhere. Such an understanding is necessary for developing an axiomatic theory but it is not practically applicable. Applications invariably discuss only distributions of probabilities of process ξ_i at some moments $t_1, t_2, ..., t_n$. Two cases are possible with time t taking discrete values (observations are made at discrete moments of time) or continuous values on some interval.

The concept of stochastic process with continuous time demands to be very cautiously treated. When understanding the relevant mathematical theorems too seriously, the realizations often acquire paradoxical properties able to direct the researcher's mind along a wrong route. I [i] mentioned the paradox concerned with the property

of the mathematical model of the Brownian motion allowing to determine precisely the coefficient of diffusion given observations of a however small interval of a realization of that motion. He who believes that this is indeed true for a physical Brownian motion will be wrong.

Here is another such example. Any broadcasting station is transmitting over a waveband of restricted width. If a radio signal is considered as a stochastic process, its spectrum will be contained in that finite interval. And there exists a mathematical theorem stating that with probability 1 a realization ξ_i of such a process is an analytical function of t. Consequently, after listening for any however short interval of time, we may unambiguously establish what was and what will be broadcast, an obviously absurd conclusion.

It is certainly easy to indicate the mistake here. First, a broadcast is not a stochastic process since it is not an element of some statistical ensemble; second, an analytical function can be reconstructed given its values on any interval only if they are given absolutely precisely which is impossible for a function of a continuous variable. Even a single number can not be written down precisely, much less a totality of an infinitely many numbers. Third, a radio signal is not an analytical function of time because in the 19th century there were no broadcasting stations whereas an analytical function vanishing on some interval vanishes everywhere.

A digression about the concept of function in mathematics is in order here[7]. At the emergence of mathematical analysis it was usually understood as a formula determining a dependence $y = y(x)$. And all functions except at a few points were continuous and differentiable. The problem concerning the proof of differentiability did not even exist. Later, however, in the 19th century an idea was established that a function is simply a relation between the sets of values of the argument x and the function $y = y(x)$. It is usually demanded that exactly one value of y corresponded to each value of x, but that the inverse was not necessarily true. And there was no cause for an arbitrary correspondence $y = y(x)$ to be continuous or differentiable.

It is rather difficult but therefore interesting to provide an example of a continuous but nowhere differentiable function. The first such example was due to Weierstrass, later other and more simple examples were discovered. Such objects proved very interesting for mathematicians and to them their attention had been to a large extent swung. For us, it is especially interesting that mathematical considerations concerning the theory of stochastic processes lead to the realization of many such processes which should be recognized as continuous but not differentiable functions (or functions only twice, say, differentiable with a continuous but not anymore differentiable second derivative).

This was indeed joyful because it apparently proved that non-differentiable functions indeed existed in nature. However, we wish to cast a shadow on that joy: it is absolutely absurd to believe that such a function can be experimentally observed. Such a realization of a stochastic process can not be given either by a formula, or a table, a graph, or an algorithm of calculation. When considering it indeed real, exactly known at all of its points, we will be able to come to absurd

conclusions. The same concerns realizations of such processes which should be analytic functions. Here, we will discuss stochastic processes with discrete time which do not tacitly contain such paradoxes as processes with continuous time and in general we may usually state that the observed values ξ_i are precisely known.

One remark concerning terminology. In Russian literature, the term *time series* usually denotes a stochastic (and often, a stationary stochastic) process, see its definition in Chapter 1. In the English literature, however, the same term denotes the values of any variable including non-random ones depending on time and observed at its discrete moments. Here, we call such objects *observational series* and will not apply the term *time series* but rather either *stochastic process* (when randomness is supposed to exist) or *observational series* (when it can exist or not). Because of causes described in Chapter 1, stationary stochastic processes are playing the main part.

The concept of stochastic process allows us to imagine a joint distribution of random variables ξ_i although usually the discussion is only restricted to bivariate distributions, and only the expectation and the correlation function

$$m(t) = \mathrm{E}\xi_i,\ B(s,\ t) = \mathrm{E}[(\xi_s - m(s)]\ [(\xi_t - m(t)]$$

are studied. For stationary processes distributions of probabilities do not change in time, so

$$m(t) = \mathrm{E}\xi_i = m \tag{3.3}$$

does not depend on t and the correlation function only depends on the difference of the arguments, $(t - s)$:

$$B(s,\ t) = B(t - s). \tag{3.4}$$

A process only satisfying conditions (3.3) and (3.4) is called *stationary in the wide sense*. Exactly this is the main concept with which modern mathematical statistics is advising to approach observational series. The theory of stochastic processes only dealing with mean value and correlation function is called *correlation theory*. We will consider it now.

3.3. Correlation and spectral theories. The main achievement of the general theory of stationary stochastic processes is the theorem establishing that in the general case the correlation function can be represented as

$$B(s,t) = B(t-s) = \int_{-\pi}^{\pi} \cos\lambda(t-s)dF(\lambda) \tag{3.5}$$

where $F(\lambda)$ is a restricted non-decreasing function. It is usually presumed that there exists a *spectral density*, i. e., such function $f(x) \geq 0$ that

$dF(\lambda) = f(\lambda)d\lambda$, so that $B(t-s) = \int_{-\pi}^{\pi} \cos\lambda(t-s) f(\lambda)d\lambda$.

Spectral analysis, that is, an experimental determination of the spectral density $f(\lambda)$, is therefore sometimes explained as the determination of the variances of the separate random components of the process. For practically applying the correlation or spectral theory it is necessary, first, to find out the practical conclusions possible from the correlation function or spectrum (spectral density); and, second, to be able to estimate the correlation function (or spectral density) by observations.

That correlation function is normally applied in statistical problems. For example, the variance of the arithmetic mean $\overline{\xi}$ is expressed through the sum of paired covariations, i. e., through a correlation function. It can also be expressed through the spectral density.

However, an estimate of a spectrum, or of a correlation function, is sometimes applied as a magic remedy allegedly making it possible to penetrate the essence of the observed process. It should be clearly imagined that the correlation theory generally deals with such characteristics that are far from determining the process as a whole and often only provides a superficial information about it. If we are interested in some problem of its structure, we must be able to formulate it in terms of the correlation theory while bearing in mind that usually we do not know precisely either the correlation function or the spectral density but estimate them by observations. We should thus consider comparatively rough characteristics determinable by issuing from non-precise data.

For example, there exists the so-called *method of canonical expansion* whose application demands the knowledge of the eigenfunctions of an integral equation in which a correlation function of a process is included as a series. This method ought to be recognized as practically hopeless because the inaccuracy of the equation's kernel very essentially influences the eigenfunctions. I do not know about any practical application of that method. All so-called *applications* issue from arbitrarily given correlation functions and do not deal with statistical material.

The estimation of the correlation function and spectrum is rather complicated. At first you should estimate and subtract the mean value m of the process. Its estimate is the arithmetic mean $\hat{m} = \overline{\xi}$. The estimate of

$$B(u) = \mathrm{E}\xi_t \xi_{t+n} - m^2$$

will be

$$\hat{B}(u) = \frac{1}{n-u} \sum_{t=1}^{n-u} \xi_t \xi_{t+u} - \hat{m}^2,\ u = 0,\ 1,\ ...,\ n-1.$$

It possesses a number of unpleasant properties. First, for an ergodic process the actual values of $B(u)$ rapidly decrease with an increase of u. However, the standard deviations of the estimates $\hat{B}(u)$ are roughly the same for any u and have order $1/\sqrt{n-u}$. Thus, for u of the order of a few dozen the magnitudes $B(u)$ themselves are very small, only hundredth and thousandth parts of $B(0)$ whereas the standard deviations (if n is not too large), tenth parts of $B(0)$ so that the estimate is senseless.

Second, these estimates $\hat{B}(u)$ when the values u are close to each other are not scattered chaotically near the real values because the neighbouring estimates $\hat{B}(u)$, $\hat{B}(u+1)$, $\hat{B}(u+2)$, ... are correlated with each other. When looking at a graph of their values the eye automatically selects rather regular oscillations, see Fig. 3, at unreasonably large values of u where actually $B(u)$ can not be distinguished from zero. Therefore, when estimating the correlation function we can not trust our eyes and all our actions become uncertain.

The estimation of the spectral density $f(\lambda)$ is preferable. When estimating it at points $\lambda = \lambda_1, \lambda_2, ..., \lambda_m$ not too close to each other, the respective estimates $\hat{f}(\lambda_i)$ will be almost independent random variables, a fact first discovered by Slutsky. For estimating the spectral density we apply the same periodogram only suitably normed. It is however very indent because its variance does not tend to vanish as the number of observations increases. Therefore the periodogram is smoothed, i. e. a mean value with some weight is taken[8] and we obtain an estimate not of the spectral density itself but of the function resulting from taking its mean with the same weight. This means that the interval of taking the mean should be small. However, that procedure when a small interval is chosen will little decrease the variance of the periodogram. Practical recommendations are here a result of a compromise between these contradictory demands.

I can not go into details of mathematical tricks and I ought to say that textbooks on the theory of stochastic processes do not usually describe the estimation of the correlation function or spectral density in any scientific manner. As I noted above, textbooks prefer to issue from a stochastic process given along with its correlation function.

As a very reliable source of information concerning statistical problems, I can cite Hannan (1960). This book is, however, very concise and difficult to read. Jenkins & Watts [1971 – 1972] is easier to read, but less reliable. For example, they do not say sufficiently clearly that none of the provided formulas for the variances of the estimates of the correlation function and spectral density is at all applicable to each stationary process; some strong conditions mathematically expressing the property of ergodicity are necessary. Nevertheless, that book is usable although regrettably their *practical examples* should be studied very critically.

I wish to warn the reader who will study the sources indicated that the initial material on which the methods of the theory of stochastic processes had been developed mostly consisted of economic data,

usually very little of them. Indeed, we may trace the change of some economic indicator over decades or at best over a few centuries (as in the case of the Beveridge series). A year usually means one observation (otherwise seasonal periodicity which we should somehow deal with will interfere, and in general most economic indicators are calculated on a yearly basis). We therefore have tens or hundreds of observations whereas calculations show that for a reliable estimate of the correlation function or spectral density we need thousands and tens of thousands of them. Already Kendall (1946) formulated this conclusion in respect of the former.

As a result, mathematicians attempt to attain something by selecting an optimal method of smoothing periodograms, but with a small number of observations this method is generally hopeless. The real applicability of the theory of stochastic processes is in the sphere where any number of observations is available. Radio physicists have long ago developed methods allowing easily and simply to obtain estimates of the spectrum of a stochastic process if unnecessary to economize on the number of observations. They apply systems of filters separating bands of frequencies (Monin & Jaglom 1967, pt. 2).

3.4. A survey of practical applications. Among the creators of the theory of stochastic processes who had also dealt with statistical materials we should mention Yale, Slutsky and M. G. Kendall (and most important are Kolmogorov's contributions, see below). Those works had appeared even before World War II, that is, when automatic means of treating the material were unavailable, and these pioneers had to work with hundreds of observations at the most.

Thousands and tens of thousands are needed in the correlation theory because we are attempting to find out too much, not an estimate of one or a few parameters, but infinitely many magnitudes $B(u)$, $u = 0, 1, 2, \ldots$ i. e. the correlation function (or spectral density, the function $f(\lambda)$ for λ taking values on $[-\pi, \pi]$).

We can choose another approach for achieving practically effective methods of correlation theory when having a small number of observations, namely, looking for models depending on a small number of parameters. Slutsky provided one such model, the model of moving average. Imagine an infinite sequence of independent identically distributed random variables

$$\ldots \xi_{-1}, \xi_0, \xi_1, \ldots, \xi_n, \ldots$$

instead of which we observe the sequence

$$\ldots \varsigma_{-1}, \varsigma_0, \varsigma_1, \ldots, \varsigma_n, \ldots \tag{3.6}$$

where

$$\varsigma_n = \sum_{k=0}^{m} \alpha_k \xi_{n-k}. \tag{3.7}$$

In other words, ς_n is a sum of some number of independent magnitudes ξ_{n-k} multiplied by suitable α_k. Slutsky modelled the system

$\{\xi_n\}$; for obtaining ς_n he superimposed a frame with a window through which ξ_n, ξ_{n-1}, ..., ξ_{n-k} were seen. For obtaining ξ_{n+1} the frame was moved one step to the right, hence the term, *moving average*. Numbers α_0, ..., α_m were parameters.

He showed that his model could provide a picture of wavy oscillations very similar to oscillations of economic indicators. However, he did not state that all the statistical properties of some observational series taken from practice are thus described. As far as I know, no such examples are provided in careful statistical works.

It is necessary to say here that statistical work with observational series demands versatile statistical checks of the adopted model. Slutsky, as well as the representatives of the serious English school such as Yule and Kendall[9] understood it perfectly well but this attitude is now regrettably lost, certainly if having in mind an average work on applications of stochastic processes.

A conviction that these processes must be universally applicable is characteristic for the bulk of publications and, as a result, the main premises with the most important of them, that the phenomenon itself should be of a statistical rather than of just an indeterminate essence, are not checked at all. A current of publications thus appears which do not deserve to be seriously considered at all. It is a fact that only a few works are left for being seriously analyzed.

Among these latter we mention first of all Yule (1927). He studies the change of the number of solar spots in time. First of all Yule rejects the model of periodogram because in that case randomness is only inherent in the errors of our measurements and does not at all influence the course of the process itself. He remarks that we ought to have some such model in which a random interference influences the subsequent behaviour of the process.

Imagine for example that we observe the oscillation of a pendulum but that naughty boys have begun to shoot it with peas. Each random hit changes its velocity and therefore influences the entire subsequent process. It is difficult to expect here statistically homogeneous shooting, but in real processes, such as solar activity or economic life statistical homogeneity of random interference sometimes possibly exists.

Let us observe the position of the pendulum at discrete moments of time (but sufficiently often, so that many observations will occur during one period of the initial oscillations). We will obtain a sequence of observations

$$\xi_0, \xi_0, ..., \xi_n, ...$$

and Yule supposes that it can be described by a model of the type

$$\xi_n + a\xi_{n-1} + b\xi_{n-2} = \delta_n \qquad (3.8)$$

where a and b are numbers (parameters of the model), $\{\delta_n\}$, a sequence of identically distributed independent random variables such that δ_n does not depend on ξ_{n-1}, ξ_{n-2}, ..., $E\delta_n = 0$ and $\sigma^2 = var\delta_n$ is the third parameter of the model.

This is the celebrated model of autoregression (of the second order) which was applied by many statisticians deserving complete trust. Yule's considerations leading to model (3.8) were, however, not quite clear. In particular, for the case of the pendulum, $\{\delta_n\}$ is not a sequence of independent random variables but is rather describable by Slutsky's moving average. However, introducing additional parameters of that average into the model will mean having too many parameters and extremely complicated work in its application.

Yule's mistake certainly does not logically prove that the model is not applicable to sunspots or some economic indicator, but of course it is a bad omen.

Descartes noted that the world can be explained in many different manners and the problem only is, to choose that which is really valid. Most chances to be valid certainly has that manner which is the most natural and harmonious and does not contain contradictions. If, however, it occurs that the creator of a theory committed a mistake at the very outset, even if only concerning a particular case, our chances of success in other cases will sharply diminish.

As to sunspots, Yule himself did not achieve a decisive positive result. He was compelled to change his model (3.8) by assuming that we observe not the variables $\{\xi_n\}$ themselves, but that our observations were corrupted by an additional random error. He had to make this change because his model did not pass a statistical check to which he subjected it, as was supposed to be done. The change of the model allows to make ends meet but in statistics introducing an additional parameter is very bad.

In general, Yule's contribution (1927) is an example of a statistical masterpiece which, however, provided a dubious (if not negative) result often happening exactly with masterpieces.

The interest emerged in forecasting stochastic processes led another representative of the English school, Moran (1954), to study the possibilities of applying model (3.8) for predicting solar activity. Since δ_n does not depend on the previous behaviour of the process, that is, on variables ξ_{n-1}, ξ_{n-2}, ..., the best possible method of forecasting the estimate of ξ_n from all the previous information is to assume that

$$\hat{\xi}_n = -(a\xi_{n-1} + b\xi_{n-2}).$$

Moran did that and had showed his result to his friends among radio physicists who told him that a forecast of such a quality could have been possible without any science, just by naked eye. And so it was, as proved by an experiment. That was the second failure of the model of autoregression.

That model possesses, however, an excellent property: it is easily applied. Its parameters are easy to estimate , the correlation function is of the kind of fading sinusoidal oscillations and is comparatively easy to be interpreted. The spectral density is also expressed in a simple way. It made sense therefore to test it many times on differing material and hope that cases in which it works well enough will be found. It is best to read about the application of the autoregression model in Kendall & Stuart (1968).

Kendall did that even before Moran's work (1954) appeared. He restricted his attention to such values of the parameters *a* and *b* in formula (3.8) which determine a stationary process, and he mostly worked with series from economics. Such series rarely oscillate around one level creating a stationary process. They usually have a tendency, a *trend*. The production of electrical energy, say, increases exponentially and therefore has a linear trend when described on a logarithmic scale. The problem consists in describing the deviations during different years from the general tendency.

Kendall thought it possible to determine the trend by some method (but certainly not by naked eye which is too subjective for a rigorous statistical school) and to subtract it. This additionally complicates the statistical structure of the remaining deviations, but there is nothing to be done about it. Exactly such deviations as though forming a stationary process were studied by the method of autoregression.

It is difficult to pronounce a definite opinion about his results. In some cases the statistical tests were happily passed, but not in other cases. May we consider that success was really achieved in those former or should we explain it only by the small number of observations? And no explanation is known why, for example, the model of autoregression with the trend being eliminated does not suit the series of the cost of wheat but suits the total head of sheep. No decisive success in treating economic series was thus achieved.

Kendall (1946) investigated the process of autoregression constructed according to equation (3.8) by means of tables of random numbers; the longest of the modelled series had 480 terms. In concluding, let us have a look at the empirical estimate of a correlation function (Fig. 3, dotted line). See how much the estimate differs from the real values (continuous line) and fades considerably slower than the real function.

Hannan (1960) published an estimate of the spectral density of Kendall's series. The graphs of the theoretical density and its various estimates are shown on Fig. 4. It is seen that they are pretty little similar to the true density. In particular, the later takes a maximal value near point $\lambda = \pi/5$ whereas the maximal values of all the estimates are at point $\lambda = \pi/15$.

An unaccustomed eye can imagine that small values of the spectral density are estimated well enough, but nothing of the sort is really taking place. The relative error is here just as great as in the left side of the graph, i. e., as for large values of the density. We see that the correlation theory, created by the founders of the theory of stochastic processes for treating discrete observational series, such as the number of sunspots in various years or the values of economic indicators exactly in those cases did not attain undoubted success.

The idea of a mathematical description of wavy processes encountered the practical difficulty in that any proper estimation of the correlation function demands not tens or hundreds of separate observations, but (Kendall 1946) tens and hundreds of pertinent waves which means thousands and tens of thousands observations. On the other hand, parametric models such as the model of autoregression had not been convincingly statistically confirmed. Consequently, the

applications of the theory of stochastic processes to that material, and to forecasting in particular, are not sufficiently scientifically justified. The worst circumstance is that many contributions are published in that field such as Ivakhnenko & Lapa (1971) which do not sufficiently check the adopted model statistically and therefore can not be considered seriously.

The situation would have been quite bad but at the same time new fields of application of the correlation theory in aero-hydrodynamics and physics which constitute the real worth of that theory were created. We will indeed consider these applications.

3.5. Processes with stationary increments. When having some mathematical tool and wishing to describe natural phenomena by its means, the most important consideration is, not to ask nature for too much, not to attempt to apply that tool in cases in which it is helpless. Thus, when imagining some wavy phenomenon, we would have liked to apply the theory of stationary stochastic processes for describing it. However, it was gradually understood that the largest waves in the observed process can either be not of a statistical essence at all, or that our observations contain insufficient data for determining their statistical characteristics, or, finally, that a purely statistical description can be short of our aims.

For example, the cyclic recurrence of economic life apparently has all these indications. Here, we can not on principle consider a phenomenon as statistical because only one realization and no statistical ensemble is available. And of course we usually have insufficient observations. Finally, a statistical description does not satisfy us because we need to know, for example, not how one or another decline or rise is developing in the mean but what happens with the particular decline or rise existing this moment.

It is absolutely impossible to reckon on describing phenomena of the largest scale in the boundaries of the theory of stochastic processes. The situation is different for phenomena on a small scale; in such cases perhaps something can be done. Take another example, the course of meteorological processes. It is absolutely clear that a statistical description of the largest changes of the weather on a secular scale is impossible and senseless. It is uncertain beforehand whether statistical methods can be applied for describing changes of the weather on a small scale during a few days, for predicting it, say. However, experience shows that this is sufficiently useless. Still, when restricting forecasts to small territories and short intervals, the success of statistical methods is brilliant. The relevant theory is called *statistical Kolmogorov – Obukhov theory of turbulence* and we will later say a few words about it.

We turn now to geology and formulate, for example, the problem of estimating the reserves of a deposit given the per cent of the useful component in a number of sample points. Here also we encounter the risk of applying stationary processes for describing that per cent over the entire deposit. The situation with the ensemble of realizations and the availability of data is very bad for determining the statistics of the largest fluctuations. On the other hand, the largest irregularities occur on a large scale and likely change smoothly; it may be therefore

expected that we know them accurately enough and do not need any statistical description. But what should be done with irregularities on a small scale which can influence the estimation of the reserves as well?

Take finally radio physics in which the concept of stationary process is recognized best of all. All kinds of interferences and noises are here usually considered as stationary stochastic processes. However, there is a special noise, the flicker noise or shimmering explained by chaotic variations of the emissive capability of the cathode electronic tubes. It is sufficiently clearly indicated, see for example Rytov (1966), that the shimmering can hardly be described by the model of stationary stochastic process.

It follows that at present we begin to realize that a mathematical description of the largest waves of wavy processes by methods of mathematical statistics is in most cases impossible. We have to reckon on describing phenomena on a smaller scale but we certainly have to forfeit much. Thus, the theory of the microstructure of turbulence is useless for predicting the weather because it does not describe the most essential phenomena occurring on a large scale. However, it is useful in other fields, for example when calculating the passage of light through the atmosphere which is important for astronomy (for taking into account the corruption of images in telescopes).

Kolmogorov introduced a universal concept of *process with stationary increments* which can hopefully replace the concept of stationary stochastic process in all the cases considered above. For discrete time it means that we turn from an observed process

$$\ldots \xi_{-1}, \xi_0, \xi_1, \ldots, \xi_n, \ldots$$

to differences

$$\ldots \eta_{-1} = \xi_{-1} - \xi_{-2}, \eta_0 = \xi_0 - \xi_{-1}, \eta_1 = \xi_1 - \xi_0, \ldots$$

and consider them a realization of a stationary stochastic process.

For processes with continuous time we turn instead from $\xi(t)$ to the derivative

$$\eta(t) = \xi'(t)$$

and call it stationary stochastic process; the differentiation should sometimes be understood in a generalized sense.

Let us explain in more detail what do we expect when turning to differences or derivatives. Imagine that the observed process is a sum

$$\xi(t) = a(t) + \varsigma(t)$$

of some random or not component $a(t)$ similar to large waves and the other component changing much more rapidly and can reasonably be called a stationary stochastic process. We recognize our inability to describe the changes of $a(t)$ and wish to study the changes on a small scale mostly determined by the other component. This is indeed

achieved by differentiating because the large component $a(t)$ likely changes slowly, so that its derivation is small. We have

$$\eta(t) = \xi'(t) = a'(t) + \varsigma'(t) \approx \varsigma'(t)$$

which means that $\xi'(t)$ practically does not include any component connected with $a(t)$. The same is achieved by taking the differences in case of discrete time.

For continuous time, rather than differentiating, we certainly can also study differences

$$\Delta_\tau \xi(t) = \xi(t + \tau) - \xi(t) \approx \int_t^{t+\tau} \eta(s) ds$$

where $\eta(s)$ is a stationary process. The second equality is needed for constructing a correlation and spectral theory of processes with stationary increments being integrals of a stationary process.

Instead of a correlation function a structural function introduced by Kolmogorov is being used:

$$D(\tau) = \mathrm{E}[\Delta_\tau \xi(t)]^2,$$

that is, the variance of the increment of the process during time τ. Practical application of processes with stationary increments can be studied by means of Monin & Jaglom (1967, pt. 2/1975).

The situation that emerged nowadays in science can be therefore described in the following way. We do not expect that general statistical methods can characterize wavy processes as a whole, i. e., including large waves. In general, the notion of stationary stochastic process is compromised. For applications, it is the turn of the concept of stochastic process with stationary increments that does not claim to cover a phenomenon as a whole but can cover it in the sphere of the small scale. Its possibilities are not yet sufficiently investigated. The situation concerning the examples with which we have dealt is this.

In economics, there exist works of the American school founded by Box, for example Box, Jenkins & Bacon (1967); Box & Jenkins (1970). However, the quality of statistical approach is there doubtful: no statistical checks are made, attention is concentrated on forecasting whereas the exclusion of the large-scale component compels us to think that it would have been better to abandon altogether predictions since they depend in the first place on the excluded component. In general, the situation is doubtful. We will consider it later.

For meteorology, processes with stationary increments are of no special significance. The Kolmogorov – Obukhov theory of turbulence rather belongs to aero-hydrodynamics. Brilliant success is achieved there: conclusions made by the creators of the theory were experimentally confirmed. That success remains, however, the only one attained.

In geology, we have the book of Matheron (1962). The factual material included there supports in some measure the hypothesis that

the structural function of the contents of the useful component is of the type

$$D(r) = \alpha \ln r + \beta$$

where r is the distance between sample points and α and β, parameters determined by observation. However, the book has a number of inconsistencies. Thus, the logarithmic dependence is continued into the interval of small values of r which is impossible because $D(r)$ is a non-negative magnitude. Then, in some cases the subject concerns the content, in other instances, its logarithm. In addition, no statistical checks are made. But still, the factual material impresses so strongly that careful reliable studies in the same direction become desirable.

In radio physics, the scientific level is high and similar inconsistencies just can not occur. However, as far as we know, no reports about successful applying the model of process with stationary increments are in existence. Rytov (1966) only formulated a hypothesis that the phenomenon of flicker should be thus described.

In concluding, I deal in more detail with the statistical theory of turbulence and the problem of forecasting.

3.6. Statistical theory of turbulence. This theory provides a brilliant success of a purely statistical description of a phenomenon, of a highly developed and very complicated turbulence with a large number of vortical movements on differing scales. Kolmogorov and Obukhov founded the basis of the theory before 1941. Experimental confirmation of their theoretical conclusions demanded perfect measuring instruments and up to 25 years. Application of that theory to problems in propagation of electromagnetic and acoustic oscillations in the atmosphere is also being developed.

A precise knowledge of the field of velocities in a turbulent current is understandably both impossible and useless. Indeed, had we some method of calculating all the velocities at all points, their registration with sufficient precision would have alone demanded an unimaginable amount of paper or magnetic tape and work with so much information is absolutely impossible. The situation should be resolved by some version of a statistical description.

It occurred that the main suitable notions can be borrowed from the correlation theory; however, in their initial form they were insufficient. There is a scientific law stating that *ex nihilo nihil fit* which means that an application of established theories does not cover anything new.

Without going into mathematical detail, I will attempt to show exactly how does this law work in case of turbulence and what new considerations it was necessary to draw for getting the things moving. Imagine a turbulent current. Its mean velocity depends on concrete conditions (what and where is the current set into motion [...]) and it is senseless to describe it by statistical methods. However, the differences of velocity in various points of the current and in differing moments of time less depend on initial conditions and to a larger extent are determined by the properties of the liquid or gas itself. So, let us study the differences

$$u(x_1, x_2, t_1, t_2) = v(x_1, t_1) - v(x_2, t_2)$$

where $v(x, t)$ is the velocity of the liquid at point x and moment t with the point x being remote from the boundaries of the current and t sufficiently large for the stationary condition to be established.

It is natural to suppose that the turbulence is stationary in the sense that the statistical characteristics of the difference u only depend on the difference $t_1 - t_2 = \tau$. The three-dimensional variables x_1, x_2 as also the difference u itself, that is, a three-dimensional vector, still remain. We have a three-dimensional field of vectors depending on six space and two temporal variables. Its statistical properties however only depend on the difference between the latter. If stopping here and expecting to determine experimentally the statistical characteristics of such a field, the experiment will invariably fail: it is practically impossible and science finds itself in a cul-de-sac.

And this is exactly the situation in some other sciences. Random stress tensors, random strength, elasticity etc can be introduced but the advantage of these notions is zero since their statistical characteristics can not be determined. Further theoretical development of the theory of turbulence was necessary, otherwise no science would have emerged there.

First of all, in a sufficiently developed turbulence all points and all directions should have the same rights. This statement seems simple but actually is rather subtle. Indeed, we can imagine a measuring device consisting of three vectors (x, e_1, e_2) the last two of them applied to the beginning of vector x and all three fixed together. An observation consists in applying the beginning of vector x to point x_1 of the current so that its end will be at point $x_2 = x_1 + x$ and we construct the projection of the difference of velocitys $v(x_2, t) - v(x_1, t)$ on directions e_1 and e_2 which will be two random variables. In correlation theory, their correlation is considered observable. This correlation should not change when the triplet (x, e_1, e_2) is rotated anyhow as a solid body nor should it depend on point x_1. Turbulence satisfying this condition is called locally *isotropic*.

It can be shown that, given such turbulence and an incompressible liquid, all the statistical characteristics of the vector field $u(x_1, x_2, t_1, t_2)$ are expressed through characteristics of any of its components, i. e., of the projection of that field on any coordinate axis. We may consider x_1 and x_2 situated on that axis and so the problem is reduced to one random function of two one-dimensional space and two temporal variables.

The reduction to one kind of variables, either space or temporal, is possible due to the *hypothesis of freezing* which means that the turbulent *curls* are carried along the main current without change, as though they were *frozen* in the liquid. In such cases we do not have to measure turbulence in various points x_1 and x_2. We arrange the line (x_1, x_2) along the velocity of the main current, put our measuring device at point x_2 and wait for the turbulence to move from x_1 to x_2. Thus, all is reduced to temporal functions only. This hypothesis (strictly speaking, its statistical characteristics rather than the turbulence itself) was checked experimentally and fit well enough.

After reducing everything to one space or temporal function, that is, to an ordinary process with stationary increments, we may expect something. Still, for determination by experiment we need the structural function, which is too much. We need it in a parameter form $D(r)$ where r is the distance between the points where the component of the velocity is measured.

The most important considerations are here due to Kolmogorov. According to them, $D(r)$ can only depend on the viscosity of the liquid which is responsible for the dissipation, the conversion of the energy of the turbulent heterogeneities into heat (and thus *reducing* turbulence) and on the amount of energy that being adopted from the main current is gradually passed from large to small whirls (and thus *supporting* turbulence). The energy is certainly considered for a unit mass of the liquid and unit time. Therefore

$$D(r) = \varphi(r, v, \bar{\varepsilon})$$

where φ is some universal function, v and $\bar{\varepsilon}$, parameters. Viscosity v is known, and the amount of energy $\bar{\varepsilon}$ is the only parameter changing from one experiment to another.

If the distance r is sufficiently small as compared with the size of the current, for the model of isotropic turbulence to be applicable but large enough so that viscosity is not yet essential for whirls of size r, then $D(r)$ does not depend on v. In this case the consideration of similarity leads to

$$D(r) = C\bar{\varepsilon}^{2/3} r^{2/3} \qquad (3.9)$$

where C is a universal constant.

For lesser r when viscosity is essential, a formula is not found although it is known that

$$D(r) = (v\bar{\varepsilon})^{1/2} \beta\left[\frac{r}{(v^3\bar{\varepsilon})^{1/4}}\right]$$

where β is some universal function of one variable. The dependence (3.9) is called the *two thirds Kolmogorov law*.

There exist spectral analogues of all those statements concerning the structural function. These conclusions were published in 1940 – 1941 and all of them were hypothetical. Intense experimental checks had begun after the war [in 1945]. Structural functions are very similar to correlation functions so that their estimates have the same unpleasant properties and it was more convenient to carry out the check by empirically measuring the spectra. No one certainly calculated smoothed periodograms, filters were used, see Monin & Jaglom (1967).

For my part, I will just say that the measurements had confirmed everything, the two thirds law for a sufficiently wide interval of the values of r, the universality of the constant C and the universal dependence of $D(r)$ expressed through function β for small values of r.

Reasoning based on common sense and the dimensionality theorem occurred exceptionally successful although they can not be absolutely precise because some physical consideration oppose them. The entire theory is of a purely statistical essence; its aim is to cover the main features of the statistics of the studied phenomenon by issuing from rather rough considerations and to approach the possibility of an experimental check. Now let us pass to a failed example.

3.7. Statistical forecast. [...] We firmly believe in scientific predictions, for example in calculations of the future situation of the planets based on the law of universal gravitation. Actually, our *belief* is certainty and it is never deceived, although the general theory of relativity is known to introduce corrections here. Are scientific methods of forecasting stochastic processes able to provide a reasonable if not firm certainty in predicting the future?

Kolmogorov and somewhat later Wiener independently developed methods of forecasting stationary stochastic processes. In his contribution on the theory of turbulence Kolmogorov clearly states that he considers his hypotheses about the structure of turbulence very likely. It is curious to compare this with the absence in his works on the prediction of stochastic processes of any hint on the possibility of practical applications.

Both in his report (1952) and in *Cybernetics* (1969?) Wiener indicated that the theory of forecasting was practically important. In the first case he stated that he was prompted by

The problem of predicting the future position of an airplane by issuing from general statistical information on the methods of its flight and from more specific knowledge of its previous path. [...] *My work was concerned with instruments necessary for realizing the theory of predicted firing in an automatic device for shooting at the airplane*

(Translated back from Russian.)

It is known, however, that such a method of shooting was not realized, not because of calculational difficulties but first of all since the path of an airplane can not be described by a model of stationary stochastic process. There possibly is a statistical component in the airplane's manoeuvre, but how can it be isolated? The manoeuvre depends so much on the concrete conditions that we can not at all discuss the statistical homogeneity of all the routes of the flight. We can attempt to isolate the statistical elements, but this problem is too difficult for being solvable under war conditions.

In all other processes, economic, technological, meteorological, etc. we usually encounter the fact that the statistical element, even if present, does not cover the entire phenomenon. Thus, only the rapid component on the small scale can yield to statistical description. And even that fact is only scientifically established in exceptional cases, for example for the microstructure of turbulence. Another such example concerns the change of the frequency of the generator of oscillations during very short periods of time, when the action of flicker and other technological causes of the change does not have enough time for being felt (Rytov 1966).

In a great majority of cases the possibility of a statistical description of at least any single aspect of the studied phenomenon is not established with certainty. Here the causes can be either an insufficient amount of experimental material or lack of understanding the need to perform all imaginable statistical checks. In such cases a statistical forecast is not more scientific than a prediction by eye which is how it is done as a rule. The only possible advantage of the former is that it can be more precise but generally its error is large and that advantage is hardly realized.

For example, if we are interested in forecasting micro-irregularities of turbulence (for which statistical homogeneity is established), the best statistical prediction of the values $\xi(t + \tau)$ of some characteristic for moment $t + \tau$ given the values $\xi(s)$, $s \leq t$, almost does not differ from the trivial forecast of $\xi(t + \tau) = \xi(t)$. Consequently, the advantage of the statistical forecast is not evident beforehand but should be experimentally established.

I (§ 3.4) have mentioned Moran's experimental prediction of the number of sunspots that indicated the uselessness of the statistical method of forecasting. Let us approach the method of prediction recently provided by Box, Jenkins & Bacon (1967) and Box & Jenkins (1970) from the same viewpoint. The method consists of two parts. First, the differences in the available observational series should be calculated and attempted to be described by a model of a stationary process being a combination of the models of autoregression and moving average. If unsuccessful, second differences should be calculated etc.

This part of the method does not give rise to any special objections; the only reservation is that the more differences we calculate, the more information about the initial process we lose. In most cases we will be able to describe the differences of a sufficiently high order, but how do we return back from them?

The second part of the method provides an answer although mathematically it is incorrect Thus, sums of infinitely many identically distributed random variables are considered, but such series are always divergent. Nowadays mathematics does not regard divergent series as negatively as previously because generalized functions enabled to make many of them sensible, but this does not concern the indicated type of series. Worse of all, these series are formally applied in the theory of conditional expectations whereas that procedure allows to provide anything.

It seems that that second part is not applicable at all to observational series described by the model of trend with error. And no statistical tests which would have excluded that model is made. In general, all the recommendations are directed to consider only correlation functions and forget the observations themselves which radically opposes a sound statistical tradition. There is therefore no guarantee that the provided method of prediction is scientifically justified; in particular, that the error of the forecast will be situated within the calculated confidence intervals.

Forecasts by eye have absolutely the same rights, the only problem is which method results in a larger error. Experimental material for

answering that question is extremely restricted. In the sources cited above there is in essence only one example of a forecast, see Fig. 5 borrowed from Box, Jenkins & Bacon (1967). The continuous broken line shows the logarithms of the monthly receipts from the sale of plane tickets during 1949 – 1960. The dotted line is the result of a forecast made from data up to July 1957. The straight lines above and below the graph show the result of an experiment consisting in smoothing the yearly extrema by a straight line and forecasting by the eye the future results.

This is shown by continuing those two straight lines through August 1957 – 1960. The forecast almost coincided with that provided by the three authors. The extrema corresponded to July or August or to one of the winter months (maxima and minima respectively) of each year. It is impossible to repeat that experiment for other months because the data on the graph are unreadable and no table of the forecast results is provided.

It is strange that Box & Jenkins (1970) did not show the described experiment on their Fig. 9.2 (p. 308). Here, their forecast is essentially better than that made by eye, and it is closer to the actual data. However, the model, its parameters and the interval of prediction, all are the same, so how can we explain the improvement? In general, the contributions of that school do not pass an attentive analysis. Borrowing an expression from the Russian author Bulgakov, their statistics can be called a statistics of *a light-weighted type* since it is presented as universally applicable and not demanding statistical checks, and it is intended to be generally popular but it does not ensure a reliable result.

The general conclusion from all the above is that we should not especially rely on statistical methods of forecasting. For applying, and relying on them we should first of all establish whether the studied phenomenon can be described by a model of stochastic process.

Notes

1. Moran (see § 3.4) possibly was an exception. O. S.

2. The separation of the random from divine design was De Moivre's main goal, see his Dedication to Newton of the first edition of his *Doctrine of Chances* reprinted in its third edition. O. S.

3. The formal introduction of least squares was due to Legendre. The author's example of an artificial object in space certainly had nothing in common with those times. O. S.

4. In the sequel, the author applied the three curves of that figure. Their equations are of the form $c_0 + c_1t^2 + c_2t^4$. All the other Figures are sufficiently described in the main text. O. S.

5. Those magnitudes are frequencies rather than periods. O. S.

6. Following a nasty tradition, Venn did not provide an exact reference, and Fisher followed suit. Abraham Tucker (1705 – 1774) is remembered for his contribution (1768 – 1778). O. S.

7. On the history of the notion of function see Youshkevich (1977). O. S.

8. Not clear enough. O. S.

9. The author apparently had in mind Karl Pearson's generally known shortcomings. Student (Gosset) was also *serious*, but Kendall did not at all belong to the Pearson school. O. S.

Bibliography

Belova L. A., Mamikonianz L. G., Tutubalin V. N. (1965 Russian), Probability of a breakdown puncture of the insulation of the coil of turbo-generators depending on the duration of work. *Elektrichestvo*, No 4, pp. 42 – 47.

--- (1967 Russian), On statistical homogeneity of the insulation of the frame of stators of turbo-generators. *Elektrichestvo*, No. 6, pp. 40 – 46.

Box G. E. P., Jenkins G. M., Bacon D. W. (1967), Models for forecasting seasonal and non-seasonal time series. In *Spectral Analysis of Time Series*. New York, pp. 271 – 311.

Box G. E. P., Jenkins G. M. (1970), *Time Series Analysis, Forecasting and Control.* San Francisco.

Hannan E. J. (1960), *Time Series Analysis*. London.

Ivakhnenko A. G., Lapa V. G. (1971), *Predskazanie sluchainykh prozessov* (Prediction of Stochastic Processes). Kiev.

Jenkins G. M., Watts D. G. (1968), *Spectral Analysis and Its Application*. San Francisco, 1971.

Kendall M. G. (1946), *Contributions to the Study of Oscillatory Time Series*. Cambridge.

Kendall M. G., Stuart A. (1968), *The Advanced Theory of Statistics*, vol. 3. London, 1976.

Matheron G. (1962), *Traité de géostatistique appliquée*. Paris.

Monin A. S., Jaglom A. M. (1967 Russian), *Statistical Fluid Mechanics*. Cambridge, Mass., 1973 – 1975.

Moran P. A. P. (1954), Some experiments on the prediction of sunspot numbers. *J. Roy. Stat. Soc.*, vol. B16, pp. 112 – 117.

Rytov S. M. (1966), *Vvedenie v Statisticheskuiu Radiofiziku* (Introduction in Statistical Radio Physics). Moscow, 1976.

Slutsky E. E. (1927 Russian), Summation of random causes as the source of cyclic processes. *Econometrica*, vol. 5, 1937, pp. 105 – 146.

Tucker A. (1768 – 1778), *The Light of Nature Pursued*, vols 1 – 7. Published under the name Edw. Search. Abridged edition, 1807.

Wiener N. (1952), Comprehensive view of prediction theory. *Proc. Intern. Congr. Mathematicians 1950*. Cambridge, Mass., vol. 2, pp. 308 – 321.

--- (1969), *Survey of Cybernetics*. London. Possible reference; author had not provided exact source.

Youshkevich A. P. (1977), On the history of the notion of function. *Arch. Hist. Ex. Sci.*, vol. 26.

Yule G. U. (1927), On a method of investigating periodicities in disturbed series etc. *Phil. Trans. Roy. Soc.*, vol. A226, pp. 267 – 298.

III

V. N. Tutubalin

The Boundaries of Applicability (Stochastic Methods and Their Possibilities)

Granitsy Primenimosti
(veroiatnostno-statisticheskie metody i ikh vozmoznosti).
Moscow, 1977

1. Introduction

I have published two booklets [i, ii]. The first was devoted to elementary statistical methods, the second one, to somewhat more complicated methods. Their main idea was that stochastic methods (like the methods of any other science) can not be applied without examination to any problem interesting for the researcher; there exist definite boundaries of that applicability.

Rather numerous comments followed, naturally positive and negative and, as far as I know, the former prevailed. In purely scientific matters a numerical prevalence (during some short period) can mean nothing; concerning publications, it is not so. The possibility of reprinting [i] for a broader circle of readers had been discussed. However, considering that problem, I have gradually concluded that during the last five years its contents had in some specific sense, see below, become dated.

The point is certainly not that previously stochastic methods should not have been applied if the studied phenomenon was not statistically stable, but that now it became possible. This could have happened if new methods not demanding that condition were developed, but science does not advance so rapidly. However, a quite definite and provable by referring to publications shift in the viewpoint on the sphere of applications of stochastic methods had happened. It will eventually make proving such a simple circumstance as the need to restrict somehow the application of the theory of probability almost unnecessary.

Then, a rapid development of concrete statistical investigations is certainly in the spirit of our time. They are difficult, demanding almost superhuman patience and insistence, but they still emerge and are being done. In a single statistical investigation, the study of statistical stability is practically impossible (and at best only if the result is negative). However, a repeated (actually, during many years) statistical investigation accompanied by checks of the conclusions on ever new material provides them quite sufficient certainty.

More precisely, we always come to understand what we know certainly; what somewhat doubtfully; and what we do not know at all. For a publication intended for a wide circle of readers it is therefore extremely important to show how should statistical investigations be carried out from the methodical point of view so that the conclusions are sufficiently certain for being practically applied. No general mathematical results are here available, this can only be done by examples.

I have published something in that direction [i, ii] but now I would have wished to accomplish such work fuller and better. Finally, for each author the aim of publication consists not only in instructing others, but to learn something himself as well. In those booklets, I have made some rather extreme statements on the practical uselessness of certain specific methods, for example [...]. It is not difficult to question such viewpoints; concerning each definite problem it is sufficient to indicate at least one successful practical application of the discussed method. Obviously neither I, nor anyone else is acquainted with all the pertinent literature but I attempted to accomplish a sample of sorts from an infinite amount of investigations so that the partisans of one or another method could have felt offended by my extreme point of view and prove the opposite.

However, concerning the application of the Bernoulli pattern to judicial verdicts, nowadays no one will probably argue; it is generally acknowledged rubbish[1]. All the other problems are, however, quite vital. I have thus considered the publication of those statements not as final conclusions but as the beginning of a big work for better ascertaining the actual situation.

It was thought that we will have to do with a comparatively small amount of concrete material. However, this is not the only essential advantage of the described method of *sampling* as compared with a full study of the publications. It is known that scientific papers are usually too short so that reading them means decoding[2] whereas in this case all difficult questions could have been resolved by asking the authors themselves.

Of course, along with really scientific objections I have received other, insignificant letters. Usually such are reports about the results of investigations in which the correspondent did not participate but only knows about them by hearsay. In such cases, since no definite data are provided, it always remains incomprehensible whether the success was achieved owing to a correct application of the theory of probability or in spite of its wrong use which is not excluded either. For example, if the report informs about the successful work of some technical system, that could have been achieved both because of a correct estimation of the essence of random disturbances but also because the designer neglected wrong stochastic estimation and guided himself by his engineer experience which had proved sufficient.

On the whole, the desired result was however achieved: I have indeed obtained objections of a scientific kind, although a small number of them. They concerned the tail areas of distributions, forecasting stochastic processes and possibilities of a periodogram analysis. Regarding the first two items, I was able to become thus acquainted with interesting and, judging by their first results, promising studies, far, however, from being accomplished. Therefore, I should not yet reject my statement that no reliable practical application of the pertinent methods is known. In spite of all of its negative essence, it is useful in that it stresses the need to work practically in those fields.

The most remarkable and scientifically irrefutable was the objection made by Professor V. A. Timofeev concerning the application of

periodograms. It occurred that work with them can be successful for example when adjusting systems of automatic regulation for isolating specific periods of disturbances so as to suppress them. The applied technique is not stochastic but I considered it necessary to describe briefly the example provided by Timofeev (§ 2.3 below).

Then, when becoming acquainted with some statistical medical problems, I encountered an apparently promising example of application of multivariate analysis (§ 2.2 below). It is almost doubtless that such methods can also be widely applied in technology for solving various problems of reliability of machinery. However, much efforts should be made for excluding the *almost*.

I thought it useful to discuss also a problem of a more general nature: what kind of aims is it reasonable to formulate for a stochastic study? Naturally, they should not be either too particular (that would be uninteresting), or too general (unattainable), see the historical material in Chapter 1.

I am sincerely grateful to the Editor, V. I. Kovalev[3], who initiated this booklet and invariably helped me.

1. Extreme Opinions about the Theory of Probability

1.1. Laplace's *singular and very facile metaphysics.* Both in teaching and during practical work I have to encounter (although ever more rarely) delusions about the actual possibilities of stochastic methods. In an intentionally rough way they can be expressed thus. *Consider some event. We are obviously unable to say whether it occurs or not. It is therefore random, so let us study it by stochastic methods.*

If you begin to argue, a few textbooks can be cited where indeed an approximately same statement (although less roughly) is written. It follows that the theory of probability is a special science in which some essential conclusions can be made out of complete ignorance. From many viewpoints (historical, psychological, etc) it seems interesting to find out the historical roots of that delusion. In general, the study of the emergence of some approach (scientific approach in particular) is extremely difficult since it usually demands an analysis of great many sources. The theory of probability was, however, lucky in some sense.

At the turn of the 18th century a greatest scholar, Laplace, summed and essentially advanced both its general ideology and concrete results. Being extremely diligent, he left a very detailed description of his views and results in his *Théorie analytique des probabilités* (TAP). We consider it permissible to restrict our attention by analyzing this single source although a strict historian of science certainly will not approve of such a view. For his part, he will be in the right; for example, it is extremely important for the history of science to study the evolution of Laplace's own ideas and his relations with other scientists, but we are actually pursuing a narrow applied aim.

In our century of rapid development of the science of science we ought to describe our source [see Bibliography]. It is a great volume containing about 58 *lists*[4] and it is pleasant to note that also in our time

only a small number of monographs are more voluminous, so that human capability of writing great books has not changed much.

The TAP is separated into two parts utterly different in style. The first part, the *Essai*, is an Introduction and summary of the book and it obeys an indispensable condition of having no formulas. Thus, the formula

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$

is expressed by words together with the definition of the numbers π and *e*. Such phrases are certainly little adaptable for perception. However, the *Essai* also contains many materials of philosophical, general scientific and applied nature described, as I see it, in a most wonderful style[5]. Had that style not been so beautiful, we would perhaps have no need to counter, after a century and a half, attempts at applying the theory of probability universally and indiscriminately.

The *Essai* is about 12 *lists* long; the rest consists of the TAP proper where Laplace applied mathematical analysis in plenty and, for us, rather strangely. This strangeness extremely impedes the understanding of the second part of the book (whereas the same is true concerning the *Essai* owing to the complete absence there of analytical formulas). It is apparently difficult to find someone nowadays who could be able to boast about having read (and understood) the TAP proper. However, many people have read the *Essai* whereas the attempts to understand the second, mathematical part led to the creation of more rigorous (and therefore more easily understandable) methods of proving limit theorems of the theory of probability. We are here only interested in the *Essai*.

As stated above, it is a work of a rather free style. A scientist's psychology is doubtlessly such that he builds a *superstructure* above his concrete scientific results. It consists of general ideas and emotions emerging out of those results and providing new faith, will and energy. The concrete results are usually published whereas the *superstructure* remains the property of a narrow circle of students and friends[6]. Laplace, however, published both and thus, as I see it, rendered his readers an inestimable service.

In his *Essai*, not being shy of the boundaries of a purely scientific publication, Laplace carried out a wide polemic. Many scientists endured quite a lot: Pascal (pp. 70 and 110)[7] for a number of unfounded statements in his *Pensées* about the estimation of probabilities of testimonies; the author of the *Novum Organum* (Bacon, p. 113) for his inductive reasoning which led him to believe that the Earth was motionless (and thus to deny the Copernican teaching); and many others, but the great Leibniz endured the most.

Leibniz is mentioned in connection with summing the series (p. 96)

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots \qquad (1.1)$$

at point $x = 1$. However, preceding the criticism of Leibniz' procedure, Laplace describes the following case, perhaps too far-fetched to be true, but characteristic of his attitude to Leibniz. When considering the binary number system, Leibniz thought that the unit represented God, and zero, Nothing. The Supreme Being pulled all the other creatures out of Nothing just like in binary arithmetic zero is zero but all the numbers are expressed by units and zeros. *This idea so pleased Leibniz, that he told the Jesuit Grimaldi, president of the mathematical council of China, about it in the hope that this symbolic representation of creation would convert the emperor of that time (who had a particular predilection for the sciences) to Christianity*[8].

Laplace goes on: *Leibniz, always directed by a singular and very facile metaphysics*, reasoned thus: Since at $x = 1$ the particular sums of the series (1.1) alternatively become 0 and 1, we will take the expectation, i. e., 1/2, as its sum. We know now that such a method of summing is far from being stupid and may be sometimes applied, but Laplace hastens to defeat Leibniz, already compromised by the preceding story.

It is indeed remarkable that now, a century and a half later, we may rightfully say the same about Laplace: *directed by a singular and very facile metaphysics*. This does not at all touch his concrete scientific work but fully concerns his general ideas connected with concrete scientific foundation. His Essai begins thus (p. 1):

Here, I shall present, without using Analysis, the principles and general results of the Théorie, applying them to the most important questions of life, which are indeed, for the most part, only problems in probability.

So, which *most important questions of life* did Laplace think about, and how had he connected them with the aims of the theory of probability? That theory includes the central limit theorem (CLT) which establishes that under definite conditions the sum

$$S_n = \xi_1 + \ldots + \xi_n$$

of a large number of random terms ξ_i approximately follows the normal law. When measuring the deviation of the random variable S_n from its expectation $\mathrm{E}S_n$ in terms of $\sqrt{\mathrm{var}\, S_n}$, we therefore obtain values of a random variable obeying the standard normal law. Briefly it is written in the form

$$\frac{S_n - \mathrm{E}S_n}{\sqrt{\mathrm{var}\, S_n}} \to N(0,1)\,.$$

Here, N(0, 1) is the standard normal distribution (with zero expectation and unit variance). Consider now the case of $n \to \infty$. If the expectations of all the ξ_i are the same and equal a, the variance also the same and equal σ^2, and the random variables ξ_i themselves independent. Following generally known rules, we get

$$\mathrm{E}S_n = \sum_{i=1}^{n} \mathrm{E}\xi_i = na,\ \ \mathrm{var}\, S_n = \sum_{i=1}^{n} \mathrm{var}\xi_i = n\sigma^2,\ \ \sqrt{\mathrm{var}\, S_n} = \sigma\sqrt{n}.$$

For a random variable obeying the law N(0, 1) typical are absolute values of the order 1. For example, the probability of its absolute value exceeding 3 is about 0.003 (hence the *three sigma* rule): we see that the inequality

$$\frac{|\, S_n - \mathrm{E}S_n \,|}{\sqrt{\mathrm{var}\, S_n}} \le 3,\ \text{so that}\ |S_n - na| \le 3\sigma\sqrt{n}$$

is practically certain.

Let $a \neq 0$. Then na is the typical value of S_n and its random deviations do not exceed $3\sigma\sqrt{n}$, a magnitude that increases with n essentially slower than na. Given a large n, the order of the *determinate component* na exceeds that of the random deviations.

Such is the purely scientific result known (at least in some particular cases) to Laplace. Let us see now what philosophical and emotional *superstructure* did he build above it. Here is one more quotation from his *Essai* (pp. 37 – 38):

Every time that a great power, intoxicated by the love of conquest, aspires to world domination, the love of independence produces, among the threatened nations, a coalition to which that power almost always becomes a victim. [...] *It is important then, for both the stability and the prosperity of the states, that they not be extended beyond those boundaries to which they are continually restored by the action of these causes*.

This conclusion is reasonable, excellent and indeed typical for the post-Napoleon France. But then Laplace adds: *This is another result of the probability calculus*. He bears in mind that, just as the determinate component prevails over randomness, see above, so also in politics, what is destined actually happens. But was it necessary to justify that statement by the CLT? For the modern reader it is quite obvious that we can only see here a remote analogy, peculiar not for science but exactly for metaphysics, and a *singular and very facile metaphysics* at that.

A bit later Laplace (p. 38) states, again citing the theory of probability: *When a vast sea or a great distance separates* a colony from the centre of the empire, the colony will sooner or later free itself because it invariably attempts to get free. And elsewhere he (p. 123) says:

The sequence of historic events shows us the constant action of the great moral principles amidst the passions and the various interests that disturb societies in every way.

He concludes that since *the action of the great moral principles* is constant, and, as the CLT teaches us, they will in any case prevail over randomness, it is better to keep to them, otherwise you will experience bad times. That conclusion is really commendable, but from the scientific viewpoint it is obviously not better than converting the Chinese emperor to Christianity desired by Leibniz. At the end of the *Essai* (p. 123) we find the celebrated phrase:

It is remarkable that a science that began by considering games of chance should itself be raised to the rank of the most important subjects of human knowledge.

He means exactly those political *applications* of the theory of probability.

All the strangeness of metaphysics in the philosophical and emotional spheres notwithstanding, Laplace shows an amazing insight when concretely applying the probability theory. I have looked through the ATP with a special aim, to find at least one wrong definite statement. It seemed that supporting myself with a hundred and fifty years during which science has been since developing and given such strangeness of his general philosophical views, it will not be difficult to find there definite errors as well. Indeed, he considered some dubious problems *on the probability of judicial decisions* etc.

It occurred, however, that it was not at all easy to find at least one wrong statement[9]. A great many applications that he considered can be separated into three parts:

1. Obvious and absolutely unquestionable problems such as partial censuses of population or the change of the frequency of male births in Paris due to foundlings.

2. Treatment of the results of astronomical observations. It is difficult to discuss those applications since vast material ought to be studied.

3. Obviously dubious problems like the probabilities of judicial decisions. Here, however, Laplace's conclusions are so careful that purely scientific errors are simply impossible.

There is nothing to say here about the first group, but something instructive can be noted concerning the second one. There, Laplace (p. 46) quotes the result of the treatment of observations: the ratio of the masses of Jupiter and the Sun is equal to 1:1071 and states that his *probabilistic method gives odds of 1 000 000 to 1 that this result is not a hundredth in error*[10]. According to modern data, that ratio is a little more than 2% larger so that the odds are obviously wrong.

The great question here is, however, was that occasioned by a mistaken treatment of the observations or by a systematic error of those observations impossible to eliminate by any statistical treatment. I was unable to answer that question. In general, it is very easy to commit such an error, and it is relevant to remark that quite recently the mass of the Moon was corrected in its third significant digit so that the precision of modern numbers should be carefully considered. If, however, we tend to believe that the observations were treated correctly, and modern numbers are also correct, we arrive at an

instructive conclusion that the presence of systematic errors ought to be allowed for.

It is interesting to quote also Laplace's viewpoint on the problems of the probabilities of judicial decisions etc. Unlike, for instance Poisson, he (p. 120) did not overestimate their reality:

So many passions, varied interests and circumstances complicate questions about these matters that they are almost always insoluble.

In essence, Laplace considered the relevant mathematical problems as models (in the modern sense of that word) and thought that conclusions of precise calculations were invariably better than the most refined general reasoning. As an example, I take up the desired number of jurors. Laplace does not attempt to find their optimal number. His only careful recommendation (p. 80) is that, having 12 jurors, the number of votes necessary for conviction should apparently be increased from 8 to 9 since, as the solution of model problems had showed him, 8 votes do not sufficiently guarantee against mistaken convictions.

Bearing in mind the exposition below, it is important to note that Laplace readily recognized the existence of problems unsolvable by the theory of probability although (see above) *the most important questions of life,*[...] *are indeed, for the most part, only problems in probability*. In our century, the following formulations are almost equivalent:

The given problem does not belong to one or another branch of science; The given problem belongs to this branch of science but is unsolvable.

1.2. Speculative criticism of the theory of probability. We see that by the time of Laplace a somewhat contradictory situation had already formed in the theory of probability. Concrete results occurred incomparably more modest than the wide perspectives imagined by him. We ought to stress that such a situation exists elsewhere as well. Thus, it is widely believed that physics considers the most fundamental laws of nature from which the laws of other, for example biological phenomena can in principle be, or will be in the remote future derived. Biology also readily speaks, for example, about the need for learning to rule the biosphere as a whole.

It seems that the psychology of a scientist is arranged in such a way that for engaging in science a certain psychological atmosphere is absolutely necessary for attaching a certain concord and generality to concrete results which often are modest and isolated. In particular, the passing of an unfailing interest in scientific pursuits from one generation to the next one can hardly be realized without working out such a psychological arrangement.

Suppose that a school student tends to choose physics as his future profession; tell him: *All your life you will have to sit by the cyclotron and measure no one knows what*, and he will hardly become a physicist. But tell him: *You will be able to contribute to the study of the most fundamental laws of nature*, and the result will be different.

The verification of the truth of a scientific proposition by practice, in the first place concerning fundamental sciences, has a special property, namely, that it often takes more than a generation. Consequently, at least because of this the transfer of interest in science from one generation to the next one is essentially important.

On the other hand, it is also important to bring that general psychological arrangement in correspondence with the actual results. Such efforts are going on in all sciences under differing circumstances. In the theory of probability the tension of passions is somewhat stronger than, say, in mathematics as a whole: it is possibly partly connected with Laplace. He was at the source of modern probability and the literary merits of his contribution laid an excessive discrepancy between its emotional and philosophical and its concrete scientific aspects.

The too wide general hopes are characterized by the emotional shortcoming of changing into disappointment once encountering a real problem. In a purely scientific aspect it consists in that the researcher, when formulating new problems, is not sufficiently critical. As a result, efforts and material values are spent on futile attempts to solve problems whereas the impossibility of achieving this would be obvious had he been a bit more critical.

In any case, certain ideas were being developed in science concerning the sphere of application of the stochastic methods. Actually, each scientist, who carried out some applied study involving probability theory, made a certain contribution to these ideas. However, their clear formulation (brilliant also in the purely literary sense) is due to Mises (1928, p. 14). He himself also attempted to construct a peculiar mathematical foundation of the theory of probability which stirred up animated criticism and at present the generally recognized axiomatization of probability is that provided by Kolmogorov (1933/1974). Nevertheless the concept itself of practical application largely follows Mises' idea.

I remind briefly this concept of *statistical homogeneity* or *statistical ensemble* (collective). For ascertaining the principles I restrict my attention to the most simple case when an experiment can either lead to the occurrence of some event A or not. Denote by n_A the number of its occurrences in n experiments repeated under presumably the same conditions. The ratio n_A/n is called the frequency of the occurrence of event A. Even before Mises statisticians (for example Poisson who studied the probability of judicial verdicts) understood perfectly well that for the applicability of stochastic methods to study the event A the stability of the frequency n_A/n as n increases should experience ever less fluctuations and tend, in some sense, to a limit (which is indeed understood as the probability $P(A)$ of A).

Mises supplemented these ideas by a clear formulation of another property that was also intuitively perfectly well understood by statisticians. Here it is. Separate the n trials beforehand into sufficiently large totalities n_1, n_2, ..., then the respective frequencies n_A/n_1, n_A/n_2, ... should also be close to each other. The separation ought to be done by drawing on the previous information; thus, two totalities could have been trials done in summer and winter with the frequencies

n_{A1}/n_1, n_{A2}/n_2, ..., becoming known after the trials. Quite admissible and practically useful is also the separation of the trials into parts of the collected material although in this case the problem of intentional or intuitive arbitrary fit becomes acute.

The demand indicated by Mises is important. Suppose that event A is the production of defective articles whose probability $P(A)$ experiences, say, seasonal fluctuations:

$$P(A) = P_t(A) = p_0 + p_1\sin(\omega t + \varphi).$$

Here t is the moment of observation, p_0 and p_1 are some constants such that $P_t(A) \geq 0$. Suppose that $t = 1, 2, ..., n$. It is not difficult to show that, for independent results of observation at those moments the ratio n_A/n will tend to p_0 (if only $\omega \neq 2\pi$). At the same time the separation according to the seasons if the seasonal fluctuations really exist will show that Mises' demand is violated. The knowledge that such fluctuations exist can be practically very important.

Here, however, a very complicated question emerges: suppose that we did not know whether seasonal fluctuations existed. How could have we suspected that the data should be separated according to the seasons? And, on the whole, is there any general method for choosing the separate groups or should we test all possible groups? We can only say that such general method does not exist and that it is obviously senseless to test all possible groups because, whatever is the situation, a certain group can contain all the occurrences of the event A, and another one, none of them so that the equality of the frequencies will be violated as much as possible. The researcher chooses the groups intuitively or bases his choice on the available pertinent information.

Then, we wish to discuss another problem: suppose that the Mises demands are fulfilled, will that be sufficient for applying stochastic methods? In other words, are those demands not only *necessary*, but also *sufficient*? Having such a general problem, we can only discuss some versions of a mathematical theorem establishing, say, that, given that the Mises conditions are fulfilled, some proposition is true, for example the law of large numbers.

Here, however, the same question emerges: how are we to choose the groups of observations? When admitting all possible groups such a demand will be contradictory, hence can not underlie a mathematical proof. If not all possible, then it ought to be stated which groups, and this is difficult.

We see that once we only begin thinking about the simplest problem concerning the possible presence of seasonal fluctuations of the probability of producing defective articles, let alone proceed to investigate it, we conclude that available general scientific prescriptions are obviously insufficient for solving a given concrete problem. I do not know even a single exception from this rule. It does not, however, follow that no practical problem can be solved at all, see below, but I note now that in spite of all the shortcomings of that concept, it still establishes absolutely clearly that some restrictions of the sphere of the application of statistical methods are necessary.

In the purely scientific sense this conclusion is not at all new. We saw how careful was Laplace concerning those stochastic applications where indeed such carefulness was needed. Poisson, although his contribution on the probabilities of judicial verdicts was wrong on the whole[11], perfectly well understood the need to verify a number of assumptions by factual materials and performed some checks obtaining an excellent fit [i]. And in general there was likely no researcher who did not somehow choose to solve such problems where the application of the theory of probability could have proved effective.

So the discussion can only concern methodical problems (methods of teaching). What should be included in textbooks intended for beginners, or in a paper designed for being widely debated? Such considerations lead to a special kind of reasoning that I am indeed calling *speculative criticism of the theory of probability*.

A student, beginning to study a subject usually does not master any concrete material. This concerns not only students of purely mathematical specialities for which the curriculum does not envisage any such material, but also those following applied specialities who study the theory of probabilities (together with all theoretical disciplines) during their first years of learning. If, however, we consider a paper discussing problems of principle, it is addressed to people who are mostly acquainted with factual materials, although different from one of them to another. This is indeed what demands a speculative discussion of the problem.

Such discussions are based on a single principle: since the necessity of restrictions in applications of the theory of probability is acknowledged, let us see whether we are able to verify their realization in practice. It is easily established that the restrictions are generally formulated too indefinitely, and if desiring to check the conclusions rather than the restrictions, we find that an exhausting verification is here also impossible.

Pertinent examples can be seen in [i] and Tutubalin (1972). However, some contributions of Alimov have become recently known. His style is very vivid, and many quotations of his statements is desirable, but we have to choose only one (1974, p. 21):

Thus, the correctness of comparing n measurements with n independent random variables is not threatened by any experimental check. Following an established tradition, such comparisons are assumed as a basis of many branches of mathematical statistics, of the theory of Monte Carlo methods, random searching, rationalization of experiments and a number of other apparently serious disciplines. Being impossible to check experimentally, they are significantly, so to say, present at the development of systems of automatic control.

Here, Alimov bears in mind that, having one sample, it is impossible to verify either the independence of separate observations or the coincidence of their laws of distribution. In general, imagining an ensemble of many possible samples given one really observable, is for him inadmissible. Accordingly, he proposes to abandon the main

notions and methods of mathematical statistics: confidence intervals, distribution of sample characteristics, criteria of fit, consistency, unbiasedness and efficiency of estimators.

In particular, the problem of Laplace's wrong estimation of the confidence interval for the mass of Jupiter[12] should have been solved simply, although, as I see it, somewhat cruelly: engineers apply confidence intervals for avoiding responsibility to the direct customer. According to Alimov (1974, pp. 31 – 32), the sense of classical formulations of a number of results essentially differs from that attributed to them by tradition, and, after being ascertained, become simply uninteresting for an applied scientist.

The quoted paper is written very expressively and clearly. The only point which we still did not understand is why does the Mises concept or the related second Kolmogorov axiomatics[13] better correspond to the interests of that scientist than the classical set-theoretic axiomatics. In any case, the assumptions of a theory can not be logically verified. His work should possibly be understood in the following way.

The concept according to, say, Ville – Postnikov[14] provides another speculatively possible approach to applied problems whereas the traditional methods of mathematical statistics then seem absurd. Consequently, if two speculative models contradict each other, at least one of them is very doubtful. However, Alimov's text indicates no decisive grounds for such an interpretation.

Alimov's views about the classical theory of probability, at least when comparing them with Laplace's understanding, are really extreme. We do not agree with them, see Chapter 2. On the other hand, we can easily imagine factual material the acquaintance with which must only lead to such views. Now, however, we note that the methodical aims, the only ones that the *speculative* criticism of probability theory is able to pursue, seems to be although not achieved, but such whose attainment is seen in principle secured.

Planck wrote[15]:

A new scientific truth does not triumph by convincing its opponents and making them see the light but rather because its opponents eventually die and a new generation grows up that is familiar with it.

It is doubtless, at least since methodology of teaching invariably follows science, that the same happens in teaching understood in a wide sense (including propaganda of some views). The point is certainly not that critical opinions (expressed, say, in my or Alimov's contributions) change the viewpoint of the public on the problems of the theory of probability. On the contrary, those works only serve as expressions of the changed public opinion. No matter that even now many university lecturers possibly keep repeating to the students that *The theory of probability studies random events; random are such events that can either happen or not.*

Yes, public opinion had changed which is reflected in new textbooks. For example, in a recently published textbook by Borovkov (1972) there is not even a trace of Laplace's *strange and very facile metaphysics*. On the whole, it is doubtless that the rising generation

ought to learn at once the simple truth that a thorough comparison of the theory with reality is necessary for the theory of probability as for other sciences.

What, however, does such comparison consist of, and how do we search for it? Alimov believes that in the most important cases this is in general impossible. Indeed, scientifically thorough works where it is done, are rather rare. We are ending this Chapter by discussing a general pertinent problem about what can we reckon on here and provide some concrete results in the next Chapter.

1.3. *The superstition of science* and a more realistic view. Alimov's proposal to abolish a larger part of mathematical statistics is not the most severe from what can be said about science in general. Tolstoi (1910) included a whole chapter entitled *False science*. His main idea was that the *empty sciences* such as mathematics, astronomy, physics do not at all answer such main moral questions like *Why am I living and how should I live*. In addition, the contents of sciences consists of separate weakly connected fragments of knowledge which had interested, no one knows why, some small group of people. And scientists had freed themselves from work necessary for life (here, Tolstoi first of all thought about the work of peasants) and are living an unreasonable life.

It is extremely interesting to see what can be answered in our time to these accusations. Nowadays, since the power of science is ever increasing, moral problems are discussed especially intensively, see for example a review of these problems (Gulyga 1975). As to the fragmentary contents of natural sciences, this is true to some extent. Indeed, we do not dwell with an all-embracing theory covering the entire nature and issuing from common principles, but with many theories of different phenomena pertaining to physics, chemistry, biology, etc. and many extremely important things do not today yield to scientific analysis.

But does it follow that the contents of science had formed randomly, only to please the whims of some people? I will try to show that this is not only incorrect, but extremely unjust (the same concerns the statement that the scientists had freed themselves from work necessary for life). At first, I allow myself an example showing the difference between science and magic. [Cf. [i, § 1.3].]

I will now allow myself some useful for understanding the problem if remote association. Let us compare the movement of science during many centuries towards certain knowledge with another century-long movement for the development of a country's North and East, for example in Russia. The Russian peasant had been able to get acclimatized and build villages only where tilling the soil was possible (practically, along river valleys).

Just the same, science had only developed where comparatively certain knowledge was possible. As a result, when looking at a map, we see clusters of villages along the rivers with practically no inhabitants in between them. Turning to science, we see that some spheres (celestial mechanics) are well developed and more than plentifully cover practical requirements, whereas we only learn how to

solve scientifically many not less important problems from weather forecasting to prevention of the flu.

Elsewhere Tolstoi compares natural sciences with pleasures, – games, riding, skating, etc, outings, – and concludes that enjoyment should not impede the main business of life. In his time, scientists apparently yet constituted such a thin layer of the population, that the great writer had no occasion to feel the labouring principle of sciences' nature[16]. Briefly, natural sciences constitute one of the many spheres of human activity with all the thus following shortcomings and merits. Consequently, for example the criticism of the theory of probability of the *speculative* kind (cf. § 1.2) can only pursue restrictive aims. Indeed, it logically shows that the premises for applying that theory can not be verified. This, however, concerns the premises of any science; although the lack of logic undoubtedly somewhat lowers the certainty of knowledge, in many cases the conclusions of probability theory still have a quite sufficient certainty for admitting them as scientific.

Many authors including Laplace discussed how the practical applicability and certainty of those conclusions is established. His reasoning in the *Essai* is not rich in content and is reduced to stating that induction was not reliable [cf. his criticism of Bacon in § 1.1] and that analogies were still worse. In my context, the response is utmost simple: the practical verification is achieved by the work of many people and many generations; they ever again return to studying a given problem.

If several large boulders were lying on a peasant's plot, he had to bypass them when ploughing. But if his son becomes able to remove them, he will do it. Just the same, in science it is not forbidden to approach old problems by new methods and either to confirm or refute the previous results. In statistics, this means that, having a small amount of data, it is impossible to say anything in a certain way, but during a prolonged statistical investigation, with new material being ever again available, no doubts are finally left.

Alimov is in the right when asserting that, having one sample, it is not at all possible to verify whether we are dealing with independent random variables. However, the situation is sharply changed after a few new samples become available. Then, in particular, we can check the previously calculated confidence intervals.

I had occasion to encounter some people keeping to logical reasoning for whom the very concept of statistical testing of hypotheses caused a feeling of displeasure. That concept from the very beginning fixes the level of significance, i. e. some non-zero probability to reject mistakenly an actually true hypothesis. Some consider this unacceptable, but the process of cognition does not consist of a single test, and even when we reject a hypothesis, we do not, happily, pass a death sentence. If new data appear, we will test it anew.

Tolstoi would have hardly rejected the viewpoint that science is some sphere of labour not higher, not lower than any other sphere (industry, agriculture, fishing etc). To support this assumption I can cite his admission, in the same book, that in its sphere of cognition of

the material world science had indeed essentially advanced. And modern development leaves no doubt in the existence of the really true science in contrast to the *false* science.

What are the practical conclusions from the considerations above? Once we acknowledge science as a kind of active human work, it follows, on the one hand, that at each moment it is incomplete and fragmentary; indeed, active work always lacks something (or even very much). On the other hand, what also follows is universality: man will always engage in science and attempt to widen the sphere of the certainty known.

In a number of fields of application of mathematics and probability theory in particular to real phenomena the situation became abnormal since the practical possibilities of application are overestimated. In such cases it is expedient to stress the unavoidable fragmentary state of all the existing applications: in mathematics, too grand intentions can occur unattainable and their inevitable failure will create for that science an extremely undesirable blow to its prestige, a situation in which science can not normally develop.

Thus, some years ago it was thought that, had there occurred a possibility of solving great problems of linear programming covering the economics of the entire nation, economic planning should be reorganized on that foundation. It is now absolutely clear that such a problem can not be either formulated or solved at least because, given that global setting, such a notion of linear programming as *set of possible technological methods* has no sense[17]. As a result, the study of local problems for which linear programming can be effective, is not at all sufficiently developed.

Awkward and absolutely useless concepts emerge when attempting to combine global problems of linear programming with a stochastic description of the possible indeterminateness. Here also only properly isolated local problems can have sense. In general, when applying the probability theory to describe an indeterminate situation, it is extremely important to attain some unity between the extent of roughing out the reality still admissible for a stochastic model and the amount of information to be extracted from reality for determining the parameters of the model. This situation is perfectly well described by the proverb: *You can not run with the hare and hunt with the hounds*. In other words, a model that adequately describes reality in detail can demand so much information for determining its parameters, that it is impossible to collect it. And a rough model only demanding a little amount of statistical information can be unsuited for describing reality. The main demand on a researcher who practically applies the theory of probability is indeed to be able to find a way out of these difficulties.

2. Logical and Illogical Applications of the Theory of Probability

Five years ago I thought it expedient to explicate, in a popular booklet, the elements of the mathematical arsenal of probability theory. However, almost at the same time as that booklet had appeared, a sufficient number of textbooks on the theory of probability had been published with the mathematical aspect being described even more than completely. Then, a tradition begins to take shape (and

wholly dominates now the teaching of mathematical analysis and a number of other mathematical disciplines) which sharply separates the pertinent contents into mathematical and applied parts.

At the beginning of the century textbooks on the theory of probability had contained very many real examples of statistical data; in the new textbooks such examples are disappearing. A natural process of demarcating teaching mathematical theory and applications is possibly going on. Indeed, had we wished to include applications in a textbook on mathematical analysis, we would have to expound mechanics, physics, probability theory and much other material.

It is a fact, however, that the applications of mathematical analysis naturally find themselves in courses and textbooks on mechanics and physics, but that the applications of the theory of probability, while disappearing from textbooks on mathematical sciences, are not yet being inserted elsewhere. It follows that the main methods of proper work with actual data and, in particular, of how to decide whether some statistical premises are fulfilled or not, are not included anywhere.

I have therefore thought it appropriate to insert here a part of these methods. They are indeed constituting its, so to say, *didactical* part. All such methods are particular, and are described in a natural way by concrete examples. However, the inclusion of a few such examples, that seemed to me important for one or another reason, pursues in addition another and more general aim. I attempted to prove that, in spite of a possible *logical groundlessness,* a stochastic investigation can provide a practically doubtless result. Confidence intervals, criteria of significance and other statistical methods to which, in particular, Alimov objects, are serving in these examples perfectly well and allow us to make definite practical conclusions. But of course, real applications of probability theory both at the time of Laplace and nowadays are of a particular and concrete type. As to my attitude towards all-embracing global constructions, it is sufficiently expressed in Chapter 1.

2.1. On a new confirmation of the Mendelian laws. We explicate Kolmogorov's paper (1940) directly connected with the discussion of biological problems which took place then[18].

At first, some simple theoretical information. Suppose that successive repetitions of an observed event constitute a genuine statistical ensemble and its results are values of some random variable ξ. The results of n experiments are traditionally denoted

$$x_1, ..., x_n \qquad (2.1)$$

(not $\xi_1,..., \xi_n$) and $F_n(x)$ is called the *empirical distribution function*:

$$F_n(x) = \frac{\text{the number of } x_i < x \text{ among all } x_1,...,x_n}{n}. \qquad (2.2)$$

This function changes by jumps of size $1/n$ at points (2.1); for the sake of simplicity we assume that among those numbers there are no equal to each other. That function therefore depends on the random

values of (2.1) realized in the n experiments and is therefore itself random. In addition, there exists a non-random (*theoretical*) distribution function

$$F(x) = P[\xi < x] = P[x_i < x] \qquad (2.3)$$

of each result of the experiment.

Kolmogorov proved that at $n \to \infty$ the magnitude

$$\lambda = \sup \sqrt{n}\, |\, F(x) - F_n(x)\, | \qquad (2.4)$$

has some standard distribution (the Kolmogorov distribution); the supremum is taken over the values of x. This result is valid under a single assumption that $F(x)$ is continuous. Now not only the asymptotic distribution of (2.4) is known, but also its distributions at $n = 2, 3, ...$

The practical sense of the empirical distribution function $F_n(x)$ consists, first of all, in that its graph vividly represents the sample values (2.1). In a certain sense this function at sufficiently large values of n resembles the theoretical distribution function $F(x)$. [...]

There also exists another method of representation of a sample called histogram [...] Given a large number of observations, it resembles the density of distribution of random variable ξ. However, it is only expressive (and almost independent from the choice of the intervals of grouping) for the number of observations of the order of at least a few tens. The histogram is more commonly used, but in all cases I decidedly prefer to apply the empirical distribution function.

The Kolmogorov criterion based on statistics λ, see (2.4), can be applied for testing the fit of the supposed theoretical law $F(x)$ to the observational data (2.1) represented by function (2.2). However, that theoretical law ought to be precisely known. A common (but gradually being abandoned) mistake was the application of the Kolmogorov criterion for testing the hypothesis of the kind *The theoretical distribution function is normal*. Indeed, the normal law is only determined to the choice of its parameters a (the mean) and σ (mean square scatter). In the hypothesis formulated just above these parameters are not mentioned; it is assumed that they are determined by sample data, naturally through the estimators

$$\bar{x};\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus, instead of statistic (2.4), the statistic

$$\sup \sqrt{n}\, |\, F_0[\frac{x - \bar{x}}{s}] - F_n(x)\, | \qquad (2.5)$$

is meant. Here, F_0 is the standard normal law $N(0, 1)$.

Statistic (2.5) differs from (2.4) in that instead of $F(x)$ it includes F_0 which depends on (2.1), $\bar{x}$ and s and is therefore random. Typical

values of (2.5) are essentially less than those of the Kolmogorov statistic (2.4). Therefore, when applying the Kolmogorov distribution for (2.5), we will widen the boundaries of the confidence region and thus admit the hypothesis of normality more often than proper.

And so, the careful practical application of the Kolmogorov test in the most elementary (and therefore most common) situation is impossible. That criterion helps in those cases when many such examples are available which were already tested by some statistical criteria and we wish to secure a general point of view concerning their numerous applications.

Let us pass now to the essence of the problem on the *confirmation* of the Mendelian laws. Here is the classical situation. Some indication has two alleles, *A* (dominant) and *a* (recessive). Two pure lines with genotypes *AA* and *aa* are taken and compulsorily crossed. A hybrid with genotype *Aa* emerges with its phenotype corresponding to indication *A*. Then a second generation is obtained under free crossing. When admitting the hypothesis of absolute randomness of the combinations of the gametes, the probability of the occurrence of genotype *aa* is 1/4. Only individuals with genotype *aa* reveal indication *a* in their phenotype so that the probability of its occurrence is also 1/4. And so, if there will be *n* individuals in the second generation, the number of occurrences of indication *a* in the phenotype may be considered as the number of successes μ in *n* Bernoulli trials with probability of success $p = 1/4$.

This is the simplest case of the Mendelian law. Vast experimental material had been collected up to 1940 from which it was seen that in many cases such a simplest law was indeed obeyed. Essential deviations (perhaps connected with a differing survivorship of individuals of different genotypes and other causes) was also revealed.

The school of Lyssenko had been attempting to prove that that law was not working. To attain that aim, experiments were carried out, in particular by Ermolaeva (1939). They were peculiar in that the material was considered not from all the individuals of the second generation taken together, but separately for *families*. It is better to explain the meaning of that term by an example. In experiments with tomatoes a *family* is consisting of all the plants of the second generation grown in the same box. Each box is sown with seeds taken from the fruit of *exactly one* plant of the first generation. The separation into families occurs quite naturally.

However, Kolmogorov (see above) showed that Ermolaeva's most numerous series of experiments can be explained exactly by the most elementary Mendel model. Suppose that for *k* families numbering n_1, n_1, ..., n_k the number of manifested recessive alleles was μ_1, μ_2, ..., μ_k, then the classical De Moivre – Laplace theorem [proving that the binomial law tended to normality] leads to the normed magnitudes

$$\mu_i^* = \frac{\mu_i - n_i p}{\sqrt{n_i pq}},\ p = \frac{1}{4},\ q = 1 - p = \frac{3}{4}$$

having approximately the standard normal distribution $N(0, 1)$; the precision of approximation is quite sufficient for n_i of the order of

several dozen. The totality μ_i^* can thus be considered (if the Mendelian model is valid) a sample with theoretical distribution being the standard normal law.

Kolmogorov studied two most numerous series of Ermolaeva's experiments and respectively two samples (2.6) with 98 and 123 observations. [...] He obtained $\lambda = 0.82$ and 0.75. The probability of a better fit (a lesser λ) was 0.49 and – 0.37 so that those values of λ were quite satisfactory.

A purely statistical investigation thus changed the results: an alleged refutation of the Mendelian laws became their essential confirmation. Apart from the opponents of the Mendel theory Kolmogorov also mentioned the work of his followers, Enin (1939) in particular. He did not subject that paper to a detailed analysis, but indicated that the agreement with the main model of Bernoulli trials was too good (the frequencies concerning separate families deviated from $p = 1/4$ less than it should have occurred according to the main model of Bernoulli trials). A detailed analysis is instructive from many viewpoints and I am therefore providing Enin's main results.

He considers the segregation of the tomato hybrids according to differing leaves: normal and potato-like. His results are separated into two groups depending on the time of sowing the seeds of the hybrid plants in the hothouse (February or April). [...]

All the material except one observation is shown on Fig. 3. We ought to decide now what kind of statistical treatment is needed. In applied mathematical statistics the application of each given statistical test is objective, [...] but which criteria should be chosen is an essentially subjective question. The answer depends on which singularities of the data seem suspicious and the statistician more or less adequately converts this impression into statistical tests. There are no common rules, we can only discuss examples.

The matter is that in principle any given result of observations is unlikely (and in our present case of a continuous law of distribution the probability of any concrete result is simply equal to zero). Therefore, a criterion can also be found that will reject any hypothesis considered in any circumstances. We ought not to be here super-diligent and only admit criteria having a substantial sense suitable for the concrete natural scientific problem. On the other hand, if not wishing to reject some tested hypothesis, it will be usually possible to choose such criteria that will not do that. Here, we are already speaking about the honesty of the statistician.

Concerning the material presented on Fig. 3, we first turn our attention to the empirical function for the first series of observations. It is situated completely above the theoretical function and in general is quite well smoothed by some straight line (dotted on the Figure) almost parallel to the theoretical. The entire difference is some shift to the left. Since we deal with a shift (we see it perfectly well, but do not know whether it is significant or not), we ought to apply the test based on the sample mean. It is equal to – 0.64 and its variance is $1/\sqrt{11} \approx 0.30$; to remind, the tested hypothesis concerns the standard normal distribution for the values of μ^*. The deviation exceeds *two*

sigma in absolute value and is highly significant. The first series of experiments is not, strictly speaking, a confirmation of the Mendelian laws.

Let us ask ourselves now, how large should the deviation be from those laws that we ought to admit when considering this series of experiments. It is certainly possible to say at once now that the discussion is pointless when declaring that such a result compels us to doubt the presence of a statistical ensemble; or, roughly the same, to doubt the independence of the separate outcomes of the experiments.

But let us try to manage by less cruel means. Suppose that each plant reveals the recessive indication independently from others, but that the probability of success (appearance of a plant with potato-like leaves) p differs from 1/4: $p = 1/4 + \Delta p$. How large should Δp be for explaining the observed shift of the empirical distribution function? Suppose that $\Delta p = -1/40$. We thought that the magnitude (2.6) with $p = p_0 = 3/4$ and $q = q_0 = 3/4$ has a standard normal distribution; actually, this will be true for

$$\mu^* = \frac{[\mu - n(p_0 + \Delta q)]}{\sqrt{n(p_0 + \Delta p)(q_0 + \Delta q)}}.$$

When calculating the difference between μ_0^* and μ^*, we may neglect the change of the denominator so that

$$\mu_0^* \approx \mu^* + \frac{n\Delta p}{\sqrt{np_0q_0}}.$$

The magnitudes n differ in different experiments, but, according to Table 1, $np = np_0 = n/4$ mostly exceeds 50, so that $n \geq 200$. Therefore, the systematic shift is

$$\frac{n\Delta p}{\sqrt{np_0q_0}} = \frac{4\sqrt{n}}{\sqrt{3}}\Delta p \approx 30\Delta p$$

and $\Delta p = -1/40$ quite well explains the systematic shift of -0.64. An estimate by naked eye using the dotted line on Fig. 3 provides -0.58, little differing from -0.64 since the mean square deviation of the arithmetic mean is $1/\sqrt{11} \approx 0.30$.

At present, there are tables of the distribution of the statistic

$$\lambda' = \sup_x |F(x) - F_n(x)| \qquad (2.7)$$

also for finite values of n, see for example Bolshev & Smirnov (1967). For the first series of observations ($n = 11$) that statistic is 0.28. It is very moderately significant for levels higher than 20%.

Consequently, when applying this test, we are not compelled to consider that the data of the first series reject the applicability of the

Mendelian laws. It is not sufficiently clear which conclusion has more natural scientific sense: either that the data do not agree with those laws, but that the discrepancy can be understood by a slight change of p (equal to 10%); or, that somewhat reluctantly we may suppose that there is no obvious contradiction with those laws.

However, Enin provides some explanation of the possible discrepancy: the plants in the hothouse sown in February suffered from a shortage of heat and light and a considerable part of the sprouted seeds perished. Plants having a recessive indication could have well had a somewhat lower probability of survival (which should be checked by a special experiment). The final results of the first series can be considered as some modest confirmation of the Mendelian laws.

We turn now to the second series. The pertinent empirical distribution function on Fig. 3 is only badly smoothed by a straight line (according, however, to my somewhat subjective opinion). In any case, the scatter of the observations is essentially less than supposed by the standard normal distribution. The most simple way to show it by a statistical criterion is to calculate the sum of the squares of the observations. It is equal to 2.85 whereas its distribution (if the checked hypothesis is valid) is the chi-squared law with 14 degrees of freedom. As indicated by the tables of that law, that value is thus practically impossible. The value of the statistics (2.7) is 0.33; with $n = 14$ that is significant at about the 5% level.

The shift of the first series of observations was in some way reasonably explained; the second series has an insignificant shift (the sample mean is -0.21) but an essentially smaller than supposed variance. The Mendelian laws are thus obeyed more precisely than supposed which is hardly possible. The most probable statistical conclusion is that the results were tampered with deliberately or not. The corruption of normality of the distribution (the impossibility of smoothing the empirical distribution by a straight line) also indicates some defect; however, for the given number of observations this conclusion would be difficult to justify by a statistical test.

In general, as far as was possible to ascertain, the trouble is apparently that the experimental data are not provided in full. And so, it is possible to confirm the Mendelian laws while intending to refute them, and it is also possible to throw them into doubt when intending to confirm them, and all of this is revealed by a purely statistical investigation.

Here, we encountered a curious violation of the order being established in mathematical statistics. When acting strictly scientifically, statistical tests should be chosen beforehand and the experiment carried out and the verdict passed only afterwards. Actually, the tests are more often chosen by issuing from peculiarities of the material noted by naked eye. They serve for checking whether these peculiarities are statistically significant or not. However, having established in our case that useful are tests based on the sample mean and the sum of the squares of the sample values, we could have, when analyzing new similar material, strictly followed statistical science.

But then the newly appearing peculiarities of that material will have not been noticed.

What kind of peculiarities could happen? For example, on Fig. 2a and 2b a certain non-zero number of observations is shown in the region $\mu^* \leq -3$. The probability of one observation being there (assuming that the Mendelian laws are valid) is 0.0014, and, of one out of approximately a hundred (to recall, the numbers of observations were 98 and 123), about a hundred times higher; here, almost precisely so. Thus, the probability of observations appearing in that region in both series is about $0.14^2 \approx 0.02$, which means that a deviation from normality $N(0, 1)$ is significant on the level $\approx 2\%$. So, are the Mendelian laws nevertheless wrong? Well, first of all, we have chosen a test corresponding to known data; second, a perfectly reasonable attitude does not mean dogmatically following tests of significance. A reasonable answer apparently means that the bulk of observations perfectly agrees with the Mendelian laws but sharp deviations perhaps do occur. It can be supposed that a deficiency in the number of displayed recessive indications has some biological sense (if, according to a very simple explanation, there exists a connection with survivorship).

Incidentally, the above sufficiently illustrates the simple idea that truth in science is established by the work of a number of generations and is not always attained at each separate investigation.

2.2. ***No one knows the hour*** ... The ancient saying, *No one knows the hour of his death*, became somewhat shaken (certainly in the statistical rather than individual sense) after life tables have been compiled and it occurred that the probability of living up to a definite age, is subject to fluctuations (depending on the conditions of life), which are, however, not too essential. A further step towards an individual forecast based on multivariate statistical analysis is partly made and partly being made. I am describing one of the most outstanding contributions in this field, the so-called Framingham investigation (one of the pertinent publications is Truett et al 1967).

The cardiovascular diseases are known to be one of the central problems of modern medicine. They are manifested in different ways; one of the most common kind is the so-called ischemic heart disease (IHD). According to the classification adopted in the cited work, it comprises cases of myocardial infarction, coronary insufficiency, angina pectoris and deaths occasioned by disturbances of the coronary blood circulation. We know well enough that the IHD often affects people yet being in the prime of creative power which makes the problem especially acute.

There exist some rather vague ideas on the part played by the factors of modern industrialized life in the development of the IHD (little physical activity, nervous-emotional stress, irrational diet, etc) and also by the possible influence of genetic factors. These ideas are certainly extremely important but we would like to have, in addition to general (but insufficiently clear and incompletely proven) ideas some amount of scientific (i. e. trustworthy) information.

That, perhaps not covering the entire problem, would provide a reliable foundation for some practical conclusions. Important is, for

example, the problem of the objectively established risk factors. To these belong, on the one hand, portents of an illness established by modern diagnostic means (e. g. changes in electro-cardiogram), on the other hand, factors of life and behaviour (age, smoking, amount of cholesterol in the blood, etc). Since the business concerns some precisely determined factors rather than vaguely understandable *excessive tempo of modern life*, a scientific investigation of their part is in principle not unlikely.

The possible ways of the development of the IHD are little known, so the statistical method of studying it is the main method. As usual, expectations here will be chiefly based on relying that a large amount of data will be able to compensate the deficiency of information about the essence of the phenomenon (in this case, of the IHD). And since that disease develops gradually, over many years, it is desirable that the investigation covers not only a large number of people, but a very long period of time as well (if possible, their whole life).

A single examination of a large number of people presents serious difficulties; and, taking into account that people usually move several times during their lifetime, you will understand that the real difficulties are great. It ought to be also borne in mind that the relative number of *cases* (of people finally developing the IHD) is rather small, so that the population to be examined mostly consists of *non-cases* (other people). Therefore, the loss of a *non-case* by the researcher is comparatively unimportant, but losing at least one *case* is extremely undesirable. However, if we allow the loss of people (for example, occasioned by the man's move or refusal to come for the examination), we do not know whether it was a *case* or not and it should be attempted that the losses be as small as possible, so perhaps the greatest part of the entire effort is spent to attain that goal.

The examination covered practically the whole population of a small American town Framingham aged 30 – 62 years at its beginning. It is going on for more than 20 years and the cited source reports the results of the first twelve years. They are based on investigating 2187 men and 2669 women not suffering initially from the IHD. Its development during those twelve years was revealed in 258 men (11.8%) and 129 women (4.8%); it was known long ago that women suffer from IHD more rarely than men.

The connection between the risk factors measured during the first examination and the probability of the development of the IHD during the 12 following years was considered. In general, it is possible to list rather many such factors, but only seven were taken account of:

1. Age (in years). **2.** Content of cholesterol in the blood serum (*mm*/100 millilitre). **3.** Systolic blood pressure (*mm* of mercury column). **4.** Relative weight (weight expressed in per cents of man's weight relative to mean weight for appropriate sex and stature). **5.** Content of haemoglobin (*g*/100 millilitre). **6.** Smoking (0, non-smokers; 1, 2 and 3, smoking less than a packet daily, smoking a packet and more than a packet). ***7.*** Electro-cardiogram (0, normal, 1, abnormal).

Treating observations whose results depend on many factors is fraught with an absolutely general difficulty and overcoming it was possibly the main finding of the work done. The point is that the result of observation (in this case, the emergence of the IHD) is generally connected with the values of the risk factors in a barely understood way. When there are a few such factors, one or two, say, the data are usually divided into intervals according to their value; in the most simple case, into two, but this is very crude and it is better to have more.

If each factor is subdivided into several levels, all their combinations should be applied to form the appropriate groups providing the frequencies of the IHD being estimates of the probabilities. These will indeed adequately describe the data (somewhat roughly because the values of the risk factors are considered approximately).

For example, the contents of cholesterol can be considered on four levels [...], the values of the systolic blood pressure also on four levels [...]. We then arrange a two-dimensional classification [...] and obtain 16 groups with the frequency of the emergence of the IHD calculated in each of them not for all 4856 observations, but for their number in the group which is 16 times smaller in the mean. Joining men and women together will likely be thought inadmissible so that the number of observations becomes about twice smaller.

In general, a modest number of observations of the order of a hundred (when having a great many total number of observations) will be left for each frequency. But what happens if we add three more groups of different ages? And four more according to the intensity of smoking? [..] As a result, we will obtain a classification with each group containing at best one observation and cases of no observations at all are not excluded. Consequently, we will be unable to determine any probabilities. [...]

The same difficulty occurs in many technical problems concerning the reliability of machinery established by several types of checks. Suppose that the results of the checks are

$$x_1, ..., x_k \tag{2.8}$$

and we would like to derive the probability of failure-free work as a function $p(x_1, ..., x_k)$. The attempt to achieve this by multivariate analysis will be senseless.

Let us see how this problem was solved in the Framingham investigation. As far as it is possible to judge, its solution had an indisputable part, but the other part was absolutely illogical. This does not mean that it is in essence wrong, but that it possibly needs some specification. The first part can be thus expounded.

When having to do with several variables, their only well studied function is the linear function; there exists an entire pertinent science, linear algebra, which also partly studies the function of the second degree. It would therefore be expedient to represent the unknown probability $p(x_1, ..., x_k)$ by a linear function. This, however, is obviously impossible because probability changes from 0 to 1 whereas

a linear function is not restricted. We will therefore take a necessary step to further complication supposing that

$$p(x_1, ..., x_k) = f(a_0 + a_1x_1 + ... + a_kx_k)$$

where f is a function of one variable changing from 0 to 1.

There still remains the problem of choosing f; many considerations of simplicity show that most convenient is the so-called logistic function

$$f(y) = \frac{1}{1+e^{-y}}.$$

Finally, changing notation to bring it in correspondence with the cited work, we have the main hypothesis in the form

$$p(x_1,...,x_k) = \frac{1}{1-\exp[-\alpha - \sum_{i=1}^{k} \beta_i x_i]}. \qquad (2.9)$$

This function is called the *multidimensional logistic function*. We have certainly not proved that the probability sought, p, *must* be expounded by (2.9), but arrived at that function without making any logical absurdities.

After formulating the main hypothesis (2.9) the parameters of that function ought to be estimated and it is here that the authors deliberately admit a logical contradiction. They suggest the model of a multidimensional normal distribution for the results of the examination (2.8). This is obviously impossible because two of the seven factors, NNo. 6 and 7, are measured in discrete units so that normality is formally impossible. Then, it is rather strange to suppose that age is normally distributed. In general, unlike the small illogicalities of choosing a statistical test when data are already available (§ 2.1), here we see a serious corruption of logic which can only be exonerated by the result obtained (cf. the proverb: *Victors are not judged*).

More precisely, the main hypothesis consisted in that there are two many-dimensional normal totalities, one consisting of the observations of the risk factors for those who were not taken ill during the next 12 years, the second concerned those who were. A problem is formulated about the methods of distinguishing these totalities.

The classical supposition of the discriminant analysis states that the covariance matrices of both totalities are the same which leads (a rather remarkable fact!) to the expression (2.9) which we arrived at by considerations of simplicity. Nowadays this probability is understood as the posterior probability of being taken ill given that the observations provided values (2.8) of risk factors. This time, however, the authors also obtained a method of estimating the parameters of (2.9). It is illogical to the same extent as the supposition of normality.

Our own reasoning which first led us to the expression (2.9) would have led us to a quite another and more complicated in the

calculational sense method of estimation of those parameters, the method of maximal likelihood. It can also be realized and it is believed that, for the data given, both methods provide results very close to each other. It is interesting, however, to see what practical conclusions were made in the cited source. After estimating somehow the values of the parameters, we can apply formula (2.9) to find out the approximate value of the probability $\hat{p}$ of developing the IHD for each examined person during the next 12 years.

The highest probabilities were observed for men of 30 – 39 and 40 – 49 years (0.986 and 0.742 that the IHD developed) and 50 – 62 years (0.770 did not develop). For women the probabilities of developing the disease were 0.838 for ages 30 – 49 and 0.773 for ages 50 – 62 [that it did not develop?]. To a certain extent these results refute the classical saying which served as the title of this § 2.2. True, it should be borne in mind that formally these figures concern *forecasting* already happened events. Such a forecast of future events is only possible if the coefficients of function (2.9) are roughly the same in another place or time as those established by the authors. Such a supposition is probable but not yet verified.

Then, having arranged the set of values $\hat{p}$ for each examined person, they can be subdivided into several equally numerous groups (of ten people, for example) such that those with the lowest values of $\hat{p}$ are placed in the first of them, people having higher, still higher, ... values comprise the second, third, ... group. Had there been no connection between the considered linear combination of risk factors with the IHD, the number of cases of that disease in all groups would have been approximately the same, but actually the emerged picture is different in principle, see for example Table 3 borrowed from Truett et al (1967). The *expected* number of cases of the IHD was determined by summing up the probabilities $\hat{p}$ of all people in the appropriate group.

In that table, we are surprised first of all by the great difference (amounting to a few dozen times) between the sickness rate in groups of high and low risk. Second, in spite of the obvious non-normality of the distributions, the results obtained by means of a normality model agree well with the actual data. The isolation of groups of people with a higher danger of developing the IHD is thus possible by issuing from the most simple clinical examination (providing the listed above risk factors). The same conclusion can be made when considering the data represented in separate age groups.

However, it should not be thought that those results are really suitable for individual forecasts. Those can only be successful for cases of very high or very low individual risk $\hat{p}$ but for all the totality the result would have been bad. This is connected with the IHD occurring nevertheless rarely (11.8% in the mean for 12 years). Indeed, issuing from the values of $\hat{p}$ we can only forecast the disease in people with a sufficiently high $\hat{p}$ and the opposite for all the others.

When choosing the boundary of the group with the highest risk in Table 3 as the critical value of $\hat{p}$, a forecast of the disease would have been wrong in 100 – 37.5 = 62.5% of cases. And on the other hand

258 – 82 – 176 cases or 68.5% of all cases of the disease would have occurred in spite of our promise of the opposite. The problem of individual forecast is therefore far from being solved.

Let us now have a look at the estimates of the coefficients in formula (2.9) and at the possible conclusions. These estimates differ for different age groups and also for women and men. For the group of men of all ages the estimate $\hat{\alpha} = -10.8986$. Other estimates are shown in Table 4. It is seen there that the coefficients of two factors out of seven (relative weight and haemoglobin) comparatively little exceed their mean square errors in absolute value. They should be recognized as less influencing the IHD than the other five. One of those five, the age, can not be changed, and it is convenient to refer the action of the rest of them with the influence of age.

For example, a daily packet of cigarettes provides 2 points and the corresponding increment of the linear function is 0.7220 which approximately means 10 years of age. In other words, smoking a packet daily brings forward by ten years the occurrence of myocardial infraction. This figure remains approximately the same in the different age groups of men. For women, the harm of smoking is represented essentially weaker. It is not quite clear why, either the figures represent reality or the number of smoking women was small (1562 women out of 2669 did not smoke at all, and only 301 used a packet daily) and the statistics was incomplete.

It is inconvenient to compare the influence of cholesterol and blood pressure with the action of age by means of Table 4. The point is that the coefficients of the linear combination of any dimensional magnitudes are also dimensional (whereas in our case, it is demanded that the linear combination be dimensionless). The comparison of the type we have applied leads, for example, to such a result as 7*mg %*[19] of cholesterol is equivalent to a year of life. We know very well what is a year of life, but is 7*mg %* much or little?

For answering that question we ought to know how large are the fluctuations of the content of cholesterol. Or, that content be divided by its mean square deviation and thus expressed as a dimensionless magnitude. After accomplishing this procedure with all the risk factors, a comparison of their coefficients provides the following arrangement of the factors in a decreasing order of importance: age, cholesterol, smoking, blood pressure, abnormal electro-cardiogram. Weight and haemoglobin influence less. The somewhat conditional character of such norming that has only a statistical sense should be understandable.

A question such as the following naturally comes to mind: If, according to Table 4, 7*mg %* cholesterol is equivalent to the same number of years of life as 4.5*mm* of mercury column of blood pressure, then what is easier to decrease, the former by 7*mg %* or the latter by 4.5*mm* of mercury column? Or, the question is formulated about dimensionless magnitudes concerning the range of the scatter, but this is not the same. Concerning the comparative possibilities of influencing the cholesterol and blood pressure, we likely enter a scientific cul-de-sac: in all probability, those are quantitatively incomparable. In general, the dominant present style of reasoning

when formulating problems about optimal solutions concerning everything happening in life usually very soon leads to a cul-de-sac.

In addition, the cited investigation concerned a totality of men that was never attempted to be influenced by pharmacological means. And it is absolutely unclear whether its quantitative characterization will persist had it been otherwise. Most likely, not, and that any comparison of 7*mm* % cholesterol with 4.5*mm* mercury column of systolic blood pressure becomes senseless.

Thus, we can not at all attribute practical significance to the cited investigation of being able to indicate a desired way of influencing risk factors for preventing the IHD. In my opinion, that investigation has no purely applied medical significance at all. However, there exists a certain set of studies for which its role should be essential. I bear in mind the examination of various medicinal preparations. Statistical investigation is the only way to obtain trustworthy results about their efficacy.

In the most simple case two totalities of men are formed by random selection, one of them (the experimental group) is treated by a preparation, the other one (the control group) gets placebo, a harmless substance similar in appearance. Results are compared, and neither is it forbidden to compare them with general statistical information about the mean rate of the IHD in a city, country, etc. However, the so-called *placebo effect* is regrettably often revealed in modern cardiology.

It consists in that the results of both groups practically coincide whereas the mean results for a larger totality are much worse[20]. This can be rather likely explained: apart from medicines, the physician applies other means for helping patients like advising him/her about a rational way of life. During modern experiments, even the doctor does not know to which group does a given man belong. As a result, against the background of general treatment by a skilled physician, the effect of chemotherapy, if it exists, is absolutely imperceptible. On the other hand, the overwhelming majority of the population either does not visit physicians, or get treated by less skilled specialists, and the results are much worse[21].

The *placebo effect* makes it impossible to judge the real benefit of pharmacological means. It can be supposed that the point is, that the frequency of the occurrence of the IHD is not so high. When considering the extreme case by supposing that some sickness rate is 1% and that some preventing means lowers it to 0.5%, any reliable estimation of the efficacy of the preparation applied should be based on at least a thousand patients (on two thousand when counting the control group). And in that control group the number of people taken ill μ_1 will obey the Poisson distribution with parameter 10, and in the experimental group, their number μ_2 will obey the same distribution with parameter 5. The probability of a wrong result ($\mu_2 \geq \mu_1$) will be, roughly, $\Phi(-5/\sqrt{15}) = \Phi(-1.29) \approx 0.10,$ which is not so small[22].

However, had we been able to select for the experiment a group of people with probability of being taken ill of 20%, then for the same twofold decrease of the sickness rate (down to 10% due to the action of the preparation) 100 patients will much better satisfy us.

The Framingham investigation indeed indicates that in principle choosing people with a many times greater probability of being taken ill is possible by issuing from a simple medical examination. Note that we supposed that the relative efficacy of a preparation is the same for all the risk groups. If this premise seems unfounded, it was perhaps unreasonable to restrict the experiment by the group of highest risk. However, afterwards it is extremely necessary in any case to separate the examined patients into risk groups for checking whether anything is (more instructive than the placebo effect) will be found when groups of roughly the same risk are compared with each other. In other words, the results of the Fromingham investigation should become a constituent part of treating the results of examining (each or at least many) medical preparations.

The stated above method naturally applies not only to medical preparations but also to many technological means intended to heighten the reliability of the work of the machinery.

Because the Framingham experiment is so methodically important, the problem of its results being justified is raised. It seems that that study provides an example of obtaining important results as well as of a comprehensive discussion of possible doubts and objections by the authors themselves.

The authors published not only positive, but also some negative results. Thus, all the numerical estimates were based on 12 years of observing a certain population, but are they applicable to other populations? Or, to ask quite sharply, is not the observed arrangement into groups of different risks just an artefact connected with selecting a rather large number of parameters? It is indeed known that an arrangement of an already collected material according to a large number of indications can be done not badly, but that the obtained formulas quit being useful when new observations are added.

The authors attempted to apply the obtained expression for $\hat{p}$ to isolate those totalities in which new cases of the IHD must occur (after the 12 years of observation). This experiment was quite successful for men aged 30 – 39 years and women aged 30 – 49 years: 8 new cases for men from a high risk group and 10 for women; 2 and 5 for those from groups of low risk. However, for other ages the experiment proved a real failure. The possible explanation is that for groups more advanced in age such a simple medical examination made more than 12 years ago was not indicative at all.

As to the applicability of the concrete numerical values of the coefficients of the function (2.9), this question can only be solved experimentally. I do not know whether that was accomplished. Usual estimates based on the model of two normal totalities do not admit the possibility of an artefact.

Thus, when analyzing one-dimensional samples (§ 2.1), it was possible to be directly convinced in the correctness of the results simply by looking at the data represented in a form convenient for understanding. For a multivariate analysis such representation is impossible which seriously hampers statistical studies in many-dimensional spaces.

2.3. Periodogram for damping fluctuations. The publication of the two previous booklets afforded me the pleasure of a very remarkable acquaintance with Timofeev, professor at the generally known Leningrad Electro-Technical Institute. Vladimir Andreevich regrettably died 5 April 1975, but he left a few books (1960; 1973; 1975) describing the now little known world of practically effective mathematical methods. In connection with the wide development of computer mathematics, which extremely broadened the scope of practically possible calculations, the attention to simple, in particular graphical methods of calculation weakened. For example, I have only come to know what is a *Lille orthogon*[23] from Timofeev's books (it is a graphical procedure for calculating the values of a polynomial also applicable for deriving the root [?]).

Of course, a computer can accomplish this incomparably faster, but in those cases in which the polynomial appears in a technical problem (and it is possible to influence its coefficients for selecting some suitable version) a graphical solution is preferable. At the same time many such methods are non-trivial inventions (similar to the invention of machines or mechanisms) almost impossible to hit upon by oneself. These inventions were being made over centuries, but now much less interest is regrettably shown for them.

Here also the metaphor comparing the progress of science with the development of a new territory (§ 1.3) comparatively accurately describes the picture: the demand for certain products fell, and the settlements existing on their manufacturing are abandoned. In science, like in other no less serious things, much depends on whims of fashion.

It is interesting to describe Timofeev's opinion about the speculative (as I named it here) criticism of the theory of probability. It invariably states that we are unable to prove logically that the premises of the theory are feasible. Timofeev noted that in essence such reasoning always has the form of reduction *ad absurdum*, but that that method, widely used in mathematics, is not mathematical but judicial, was obligatory in courts of ancient Greece exactly at the time when geometry had been formed. Perhaps it came to mathematics from pleadings.

And now the periodogram. Periodic dependences should be isolated in a series of observations x_0, x_1, ..., x_n (or, for the continuous case, in $x(t)$, $0 \le t \le T$). To achieve this aim, some method of comparing the observations with an ideally periodic function

$$e^{i\omega t} = \cos\omega t + i\sin\omega t$$

or with some other periodic function is applied. Most complicated is the determination of a latent period (or a few periods) in our observations.

In mathematical statistics there is a pertinent method consisting of calculating the expression

$$I_T(\omega) = \left| \int_0^T e^{-i\omega t} x(t)dt \right|^2$$

or, for discrete observations,

$$I_T(\omega) = \left| \sum_{t=0}^{n} e^{-i\omega t} x_t \right|^2$$

at many values of ω and determining one or a few maxima of $I_T(\omega)$. Actually, however, as is possible to find in Timofeev's book[24], there exist a few other expressions differing from $I_T(\omega)$ by the limits of integration (summation) and also by taking into account not only the modulus, but the argument of the complex magnitudes as well.

Various graphs are thus obtained whose behaviour at different ω allows to localize the possible periods contained in the observations. Strictly speaking, Timofeev's considerations are not stochastic since a stochastic approach demands to apply the notion of an ensemble of imagined realizations of the observations (of which we see only one) and to make estimations based on these notions of the statistical significance of the isolated periods. This is possible if certain assumptions about the applied model describing the observations are made.

For example, it is very convenient if the model, in the discrete case, is of the kind

$$x_t = \sum_{k=1}^{n} a_k \sin(\omega_k t + \varphi_k) + \xi_t \qquad (2.10)$$

where $\xi_0, \xi_1, ..., \xi_n$ are *independent* identically distributed random variables. In the continuous case such a model with independence of the values of *noise* $\xi(t)$ at any no matter how close values of t (*white noise*) is less realistic. It is possible to provide a number of physical examples where model (2.10) is realistic.

The first example to come across I can mention, can be provided by the observations of the brightness of a variable star if measured not too frequently. The reasonableness of the use of the periodogram method and the possibility of certain estimates of significance in such cases is doubtless. However, in more complicated cases, in which we are unable to discuss a stochastic model of noise corrupting the process, stochastic statistical considerations with estimation of significance are impossible.

Basing myself chiefly on the application of the periodogram method to series in economics, I [ii] formulated a number of sceptical comments on the actually achieved success. It were these remarks that prompted the only scientifically doubtless objection mentioned above in the Introduction, and it came from Timofeev. He indicated a very interesting unexpected example of a practical application of periodograms. To repeat, in this case the study has no stochastic essence (isolation of some undeniable peaks on the periodogram whose significance is not needed to estimate). [The author describes that successful and important industrial case.]

Conclusion. Some Problems of the Current Development of the Theory of Probability

The examples provided in Chapter 2 were aimed at illustrating the idea that the problem concerning the boundaries of applicability of the theory of probability can not be solved speculatively, by logical justification (or by justifying the opposite). Neither does a single practical success scientifically assure us in the correctness of a theoretical concept. [...]

Only prolonged studies lasting many years (almost 20 years in the Framingham investigation (§ 2.2)) and even carried out by many generations of scientists (like the study of problems of heredity originated by Mendel) provide a reliable result. In a purely methodical sense such studies ensure complete possibility of experimental checks of many stochastic assumptions. In particular, checks of statistical homogeneity (for example, by non-parametric criteria for distinguishing two empirical distribution functions), of confidence intervals (recall my rejection of that interval for the mass of Jupiter in § 1.1) and of much more.

And so, it is wrong that no experimental checks are threatening those premises (Alimov's objection). However, if simply collecting the (statistical or not) ensemble of all the instances in which stochastic methods are applied, and find out in how many cases Alimov and I were in the right, then, as I fear, he would have collected an overwhelming majority of votes. I would have to take cover behind the argument that in science a numerical majority of votes might mean nothing.

All the circumstances concern one aspect of the problem, of what and under which conditions can theory give to practice. Let us try to think what, on the contrary, can practice give to theory. For mathematics, this is a venerable question and most extremely pertinent opinions had been voiced. I begin by quoting the opinion of the celebrated French mathematician Dieudonne (1966, p. 11; translated now from Russian):

In concluding, I would wish to stress how little does the most recent history exonerate the pious banalities of the soothsayers of a break-up who are regularly warning us about the pernicious consequences that mathematics will unavoidably attract to itself by abandoning applications to other sciences. I do not wish to say that a close contact with other fields such as theoretical physics is not beneficial for both sides. It is absolutely clear, however, that among all the astonishing achievements described, not a single one, possibly excepting the theory of distributions, is at all suitable for being applied in physics. Even in the theory of partial differential equations the emphasis is now much more on the internal and structural problems than on those having a direct physical significance.

Even if mathematics be cut off forcibly from all the other streams of human activity, it will still have food enough for centuries of thought about great problems which we still ought to solve in our own science.

What objections can be made? First, since the problem concerns an interval of a few centuries of time, it will be advisable to turn to history and look whether there are examples of what is happening with some fields of intellectual activity the interest in which is preserved as long as that. An example of such an activity is the scholasticism of the Middle Ages.

Scholastics were clever and diligent. In any case, the volume of their contributions was of the same order as, say, those of Laplace (the amount of paper that a man can cover with writing during his lifetime likely little depends on the contents of the written). Universities and academies had been initially created for scholastics because of the importance of the moral and ethical applications of their work, actual or imagined. No one had expelled them from those institutions with *a red-hot broom*[25], but it somehow happened all by itself that scientists, physicists, mathematicians, chemists etc., had occupied their places. Why did that occur?

I believe that the reason was that scholasticism had gradually withdrawn into its own business and quit to provide society solutions of moral problems essential for everyday life. For example, now, as in the Middle Ages, each solves for himself whether to marry or not. Scholastics naturally discussed that problem but their answers became long summaries of the diverse opinions of fathers of the church and ancient philosophers[26]. What then should have done a *practically working* clergyman when asked by his parishioner?

Such questions gradually occurred unbecoming to serious science and then it somehow happened all by itself that the society had begun to consider unbecoming scholasticism as a whole. This example compels us to think carefully what would have happened to mathematics had it been for centuries cut off from all the streams of activities. The action of the ensuing phenomenon would have been very simple: the number of young men wishing to devote themselves to mathematics would have gradually decreased since those other streams indeed play the most important part in attracting their interest.

However, finally it is possible to admit, and Dieudonne's article convinces us, that there exists mathematics of different types; one is directed towards its own interests, the other one, towards applications. Both have a quite lawful right to exist because, for example, the Kolmogorov axiomatics of the theory of probability, necessary in a sense for applications as well, had emerged on the basis of the theory of functions of a real variable (obviously belonging to the first type). But then, to which type does the theory of probability belong?

The distinguishing feature of mathematics of the first type is its somewhat special elegance (presenting a comparative simplicity which makes it possible to perceive that quality). The theory of probability has rather many results of exactly that kind, mostly connected with the Kolmogorov axiomatics and resembling the theory of functions of a real variable. However, the main contents of that science having been formed at the time of Laplace[27], developing after him and being elaborated nowadays is not, alas! beautiful at all. For example, limit theorems are usually rather decently formulated, but as a rule their proofs are helplessly long, difficult and entangled. Their sole raison

d'être consists in obtaining comparatively simple stochastic distributions possibly describing some real phenomena[28].

Turning to reality always refreshes whereas severing the connections with it spells danger of a scholastic degeneration. Mathematics is wonderful, but at the hands of its separate representatives it can degenerate, for example into scholasticism. It is regrettably sufficiently easy to overstep the limits beyond which begins scholasticism. Internal problems strongly attract. A man always wishes to tidy his own home both because it is his home and because it is the easiest. So where is that dangerous limit overstepping which we will only be floor polishing in our own apartment without providing anything for the society? That limit is only well seen in a historical perspective, but at each moment it is extremely indefinite and unsteady.

In a strange way statistics can partly help here, this time assuming the aspect of science of science (Nalimov et al 1969). Rather recently a comparatively simple method was applied. It consisted of a formal study of the bibliographies appended to each scientific paper. The number of those interested in a given circle of problems can be roughly estimated by perceiving which groups of authors quote each other. Attempts to build up some system of administrative estimates by basing yourself on such studies will certainly cause all the authors to cite each other in a purely formal way; and it is practically impossible to distinguish whether references were essentially needed or included as a payback.

However, without any administrative pressure the study of references is a valuable and more or less objective method of science of science. And such analysis shows that in probability theory only very small (as compared, for example, with physics) groups of authors refer to each other. This means that the interest has narrowed which was largely caused by its unwieldy mathematical machinery and which is a typical sign of a scholastic degeneration.

Perhaps the simplest way to combat that danger is to turn to physical applications. Their seriousness was never doubted by anyone, and here the interest now concentrates in particular around the problems of statistical physics. Most wide and complicated mathematical arsenal able to satisfy various tastes is applied. Physical problems are also interesting in that very much can be done by mathematical means.

However, much more variable is the field of so to say purely statistical applications. In very many important matters a far reaching mathematical analysis is impossible, but if a vast statistical material can be available, it can compensate to a necessary extent the scarcity of theoretical ideas. In all such cases statistical treatment is one of the main means of study. Here, I provided examples of exactly that kind absolutely leaving aside the doubtless problem of physical applications.

In purely statistical problems the main part is played by some stochastic model of the phenomenon. In the simplest case the supposed kind of distributions of the observations (normal, exponential, Weibull, etc) can be understood as the model. In more complicated cases the model gets more complicated as well; the theories of

reliability and queuing are known to apply rather complicated analytical models. A certain disproportion in the current development of probability theory consists in that a rather large number of theoretical models (even analytically studied to a sufficient extent) is collected, but at the same time in many cases they were never practically compared with reality. Of course, a creation of a theoretical model marks a necessary initial period without which no such comparison, and no understanding of the actual data is at all possible. However, too often a study stops at that period. At the same time, each comparison with reality usually calls into being new models, that is, acts refreshing in that sense as well.

Figures and Tables

I did not reproduce them. Fig. 1 – 3 and Tables 1 and 2 concerned the papers of Ermolaeva (1939) and Enin (1939). Tables 3 and 4 from § 3.2 explained the Framingham experiment. Table 3 provided the expected and actual number of taken ill, in each expected interval of risk, separately for men and women. Table 4 showed the estimates of the coefficients of the factors of risk.

Notes

1. See [i, Note 4]. O. S.

2. This was a feature of Soviet publications (perhaps of Russian papers even now). The late Professor Truesdell told me that he was unable to read them in translation (also because translations are usually quite formal. O. S.

3. Strangely enough, no editor is mentioned in the booklet. O. S.

4. A *list* is 24 pages typescript or 16 pages of published text. O. S.

5. No wonder Laplace was elected member of the French Academy (not to be confused with the Paris Academy whose member he also was) devoted to the study of the French language. O. S.

6. Some scientists (Chebyshev, Markov) did not have any *superstructure*. O. S.

7. The author referred (not in all necessary cases) to the text of the TAP as published in 1886. Instead, I provided references to its English translation (2005/2009). O. S.

8. This is my quotation from Laplace (2005/2009, p. 97) inserted instead of the author's description. O. S.

9. There are mistakes. One of them, noticed by Pearson, concerned his model of births and deaths, see Sheynin (1976, p. 160). Then, he had been keeping to his own practically useless theory of errors and thus caused French authors to shun Gauss (Sheynin 1977, pp. 52 – 54). Laplace's astonishing mistake (1796/1884, p. 504) was to state that the planets moved along elliptical paths not in accordance with Newton's discovery, but because of small differences in densities and in temperatures of their various parts. O. S.

10. Concerning the precision of his estimate, Laplace (2005/2009, pp. 46 – 47) stated: *after a century of new observations* [...] *examined in the same way* [...]. See also Cournot (1843, § 137). O. S.

11. Once more, see [i, Note 4]. O. S.

12. See Note 10. O. S.

13. The *classical set-theoretic axiomatization* is thus called the first one. Gnedenko (1969, p. 118), in a brief survey of the history of this problem, named only one pertinent publication, Lomnicki (1923). O. S.

14. Concerning Ville see, for example, Shafer & Vovk (2001, pp. 48 – 50). The other reference is Postnikov (1960) O. S.

15. I have found this translation in Google with a reference to a commentator of Planck. O. S.

16. Surely Tolstoi knew about such most actively working scholars as for example Mendeleev or Chebyshev. O. S.

17. The author had understandably chosen an ideologically safe cause. In 1959 Kolmogorov (Sheynin 1998, p. 542) was much more specific. It was necessary, he stated, to *express the desired optimal state of affairs in the national economy by a single indicator*. Read: to abandon the Marxian *socially necessary labour* as indicator of cost and measure cost in monetary units. O. S.

18. This is an understandably mild expression. Actually, genetics was uprooted as decided beforehand by the party's leadership, many scientists severely persecuted (Vavilov, the world renown scholar, died in prison) and even Kolmogorov's paper (1940),see below, was considered dangerous. In 1950, Gnedenko (Sheynin 1998, p. 545) mildly criticized it (undoubtedly after discussing the matter with him). In 1948, Fisher most strongly condemned Lyssenko (Ibidem, p, 544).

For the sake of comprehensiveness I add references to Lyssenko and Kolman, a high ranking party apparatchik who at the end of his life did not return from a visit to his sister in Sweden and then published a book with a telltale title. O. S.

19. This is hardly understandable. O. S.

20. What does this mean actually? O. S.

21. Is this really connected with the placebo effect? O. S.

22. No explanation provided. Notation Φ usually meant the distribution function of the normal law. O. S.

23. This is my attempt of translating that expression from Russian. I did not find it in any other language. O. S.

24. The author provided a wrong reference. O. S.

25. This was a beloved expression of the Soviet press applied in appropriate cases. O. S.

26. The initial aim of scholasticism was the study of Aristotelian philosophy but soon it turned to uniting philosophy and theology. Accordingly, the first universities consisted of three faculties devoted to theology, canon law and medicine so that scholasticism had indeed been avidly taught there. It was gradually excluded by the developing natural science although its structure proved useful for logic. One of its teaching was the so-called probabilism, see [i, Note 3].

Rabelais, in his immortal *Gargantua and Pantagruel*, had left a vivid picture of the benefits of gaining useful knowledge (rather than repeating Aristotle or Thomas Aquinas). There also the problem of a possible marriage is shown to depend on circumstances. O. S.

27. Modern probability appeared in the 1930s when such notions as density had begun to be considered as mathematical entities. O. S.

28. See [iv, Note 2]. O. S.

Bibliography

Alimov Ju. I. (1974 Ukrainian), On the application of methods of mathematical statistics to treating experimental data. *Avtomatika*, No. 2, pp. 21 – 33.

Bolshev L. N., Smirnov S. V. (1967 Russian), *Tablitsy Matematicheskoi Statistiki* (Tables of Mathematical Statistics). Moscow, 1968.

Borovkov A. A. (1972 Russian), *Wahrscheinlichkeitsrechnung. Eine Einführung*. Berlin, 1976.

Cournot A. A. (1843), *Exposition de la théorie des chances et des probabilités*. Paris, 1984.

Dieudonne J. (1966), *Sovremennoe Razvitie Matematiki* (Current Development of Mathematics). Coll. Translations. Place of publ. not provided. Original contribution not named.

Enin T. K. (1939 Russian), The results of an analysis of the assortment of hybrids of tomatoes. *Doklady Akademii Nauk SSSR*, vol. 24, No. 2, pp. 176 – 178. Also published at about the same time in a foreign language in *C. r. (Doklady) Acad. Sci. URSS*.

Ermolaeva N. I. (1939 Russian), Once more about the "pea laws". *Jarovizatsia*, No. 2, pp. 79 – 86. Note the scornful term for the Mendelian laws.

Gnedenko B. V. (1969 Russian), On Hilbert's sixth problem. In *Problemy Gilberta*. Moscow, pp. 116 – 120. German translation of book: *Die Hilbertschen Probleme*. Leipzig, 1971 (Ostwald Klassiker No. 252), see pp. 145 – 150.

Gulyga A. V. (1975 Russian), Can science be immoral? *Priroda*, No. 12, pp. 45 – 49.

Kolman E. (1939 Russian), Perversion of mathematics at the service of Mendelism. *Jarovizatsia*, No. 3, pp. 70 – 73.

--- (1940), Is it possible to prove or disprove Mendelism by mathematical and statistical methods? *C. r.* (*Doklady*) *Acad, Sci. l'URSS*, vol. 28, pp. 834 – 838.

--- (1982), *We Should Not Have Lived That Way*. New York. In Russian with an additional English title.

Kolmogorov A. N. (1933 German), *Osnovnye Poniatia Teorii Veroiatnostei* (Main Concepts of the Theory of Probability). Moscow, 1974.

--- (1940), On a new confirmation of Mendel's laws. *C. r.* (*Doklady*) *Acad. Sci. l'URSS*, vol. 28. No. 9, pp. 834 – 838.

Laplace P.-S. (1796), *Exposition du système du monde. Oeuvr. Compl.*, t. 6. Paris, 1884. Reprint of edition of 1835.

--- (1812), *Théorie analytique des probabilités*. *Oeuvr. Compl.*, t. 7. Paris, 1886.

--- (1814 French), *Philosophical Essay on Probabilities*. New York, 1995.

Lomnicki A. (1923), Nouveaux fondements du calcul des probabilités. *Fondam. Math*., Bd. 4, pp. 34 – 71.

Lyssenko T. D. (1940), In response to an article by A. N. Kolmogorov. *C. r.* (*Doklady*) *Acad. Sci. l'URSS*, vol. 28, pp. 832 – 833.

Mises R. (1928), *Wahrscheinlichkeit, Statistik und Wahrheit*. Wien.

Nalimov V. V., Mulchenko Z. M. (1969), *Naukometria* etc. (Science of Science. Study of the Development of Science As an Informational Process). Moscow.

Planck M. (1960), *Edinstvo Fizicheskoi Kartiny Mira* (Unity of the Physical Picture of the World). Coll Translations. Moscow. Original contribution not named.

Postnikov A. G. (1960), *Arifmeticheskoe Modelirovanie Sluchainykh Prozessov* (Arithmetical Modelling of Stochastic Processes). Moscow.

Shafer G., Vovk V. (2001), *Probability and Finance. It's Only a Game*. New York.

Sheynin O. (1976), Laplace's work on probability. *Arch. Hist. Ex. Sci*., vol. 16, pp. 137 – 187.

--- (1977), Laplace's theory of errors. Ibidem, vol. 17, pp. 1 – 61.

--- (1998), Statistics in the Soviet epoch. *Jahrbücher f. Nationalökonomie u. Statistik*, Bd. 217, pp. 529 – 549.

Timofeev V. A. (1960), *Teoria i Praktika Analiza Resultatov Nabliudenii* etc (Theory and Practice of Analysing the Results of Observation of Technical Objects etc). *Trudy Leningradsky Electro-Techn. Inst*.

--- (1973), *Matematicheskie Osnovy Tekhnitcheskoi Kibernetiki* (Mathematical Elements of Technical Cybernetics). Lecture notes. Pensa.

--- (1975), *Inzenernye Metody Rashcheta i Issledovanie Dinamicheskih System*. (Engineering Methods of Calculation and Study of Dynamic Systems). Leningrad.

Tolstoi L. N. (1910), *Put Zizni* (The course of life). *Poln. Sobr. Soch.* (Complete Works), vol. 45.

Truett J., Cornfield J., Kannel W. (1967), A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Diseases*, vol. 20, pp. 511 – 524.

Tutubalin V. N. (1972 Russian), *Teoria Veroiatnostei* (Theory of Probability). Moscow.

IV

Yu. I. Alimov

An Alternative to the Method of Mathematical Statistics

Alternativa Methodu Matematicheskoi Statistiki. Moscow, 1980

Introduction

Both mathematicians and those who have been applying mathematics are often recently expressing their concern that in many instances mathematical models noticeably alienate from reality. As a consequence, the work of highly qualified specialists and valuable computer time is used with insufficient effect. Criticism, occasionally very sharp, of this situation is seen ever oftener in papers and monographs for specialists and in textbooks and popular scientific editions, see for example Blekhman et al (1976); Grekova (1976); Venikov (1978); Vysotsky (1979). It is indicative that a paper of D. Schwarz called *On the pernicious influence of mathematics on science* is didactically quoted in Venikov (1978).

In particular, models offered by mathematical statistics are often remote from reality. Tutubalin's booklets [i – iii] are devoted to the conditions and boundaries of the applicability of stochastic methods, and much attention is shown to such problems in his textbook (1972). With respect to its *restrictive* direction, this booklet adjoins those publications. I stress at once that my contribution is not at all opposed to statistics as such.

I understand statistics as any calculation of means or other *combined* treatment of experimental data aiming at providing their *predictable integral characteristics*. It is assumed that these will be later measured for future similar experimental data so that the correctness of the statistical forecast will be actually checked.

I am not at all against the use of mathematics in statistics either; otherwise, the latter is simply unthinkable so that below I am treating *mathematical statistics*. Choose any pertinent treatise and you will be easily convinced that by no means any application of mathematics in statistics is understood as mathematical statistics. After attentively looking, it is seen that mathematical statistics is a very specific discipline possessing its own peculiar method whose distinctive feature is the *conjecturing of exactly one storey* of probabilities called confidence probabilities or levels of significance *above those really measured in an experiment*. It is possible to disagree with such a specific approach.

Mathematics can be applied in statistics in a manner somewhat different from what is prescribed by mathematical statistics.

In practice, the principles of statistically treating experimental data which are being applied for a long time now have nothing in common with confidence probability and are therefore alien to the foundation of mathematical statistics. We find for example that [a certain magnitude] is equal to 0.0011609 ± 0.000024. Here, only the maximal error of the

measurement is provided. Recently, physicists have sometimes begun to indicate instead the mean square error of measuring the last digits of the experimental result, usually in brackets; for example, the velocity of light in vacuum is [...]. Essential here is that unlike confidence probabilities of mathematical statistics, the maximal and the mean square error were actually measured.

For many years, mathematical statistics has been actively propagandized, but still perhaps even nowadays physicists will be unable to refrain from smiling had we told them, say, that after treating the observations of the velocity of light, c, according to the prescriptions of mathematical statistics, c is situated in such-and-such confidence interval with confidence probability $P = 0.99$ and within a more narrow interval with $P = 0.95$.

I also refer to physicists in the sequel. It was in physics that the basis of modern exact natural science had been laid, the largest amount of experience of complicated and subtle experimentation accumulated and a developed culture of a sound treatment of experimental data had been achieved. On the other hand, it was physics that provided the example of applying mathematical structures which is now often recognized not favourably enough for other fundamental and applied disciplines. I return to that problem at the end of my booklet.

Its main aim is to describe the principles of such a treatment of data that abstains from mentioning confidence probabilities. These principles had appeared even before mathematical statistics had; indeed, appeared at the same time as the first quantitative experimental results in natural science did. However, they were reflected in the theory of probability only much later during the process of the development of the approach connected with Mises. This approach has been vividly discussed for decades, see my papers and textbooks (1976, 1977, 1987b; 1978a; 1979).

The connection of that Mises approach with the principles and methods different from those of mathematical statistics is fundamental and the contents of this booklet is therefore largely reduced to a consistent although only understandably sketchy description of that approach. Such an exposition is still lacking in the literature easily read by a broad circle of readers.

I am concentrating on the problems of *interpretation and practical application* of stochastic notions. Unlike the solution of purely mathematical issues, any answers to such problems are always to a large extent arguable and the reader ought to take it into account. I am describing an approach noticeably different from that of the standard treatises and most works on probability theory and mathematical statistics and I repeat that my point of view is not at all new. Its extreme version is nicely expressed, for example, by Anscombe [1967, p. 3 note]: it is inadmissible to *identify statistics with the grotesque phenomenon generally known as mathematical statistics*.

1. Introductory Remarks about Forecasting

The final aim of research in both fundamental and applied natural science is a reliable *forecast* of the results of future experiments. By *experiment* I mean not only investigative, reconnaissance trials, but

also the operations of various devices and systems. I also understand prediction as designing all kinds of instruments, devices, systems etc. You can say that a forecast as a demand of reproducing a published result was being accepted as a definition of the final aim and distinctive feature of natural science even at their birth.

That demand apparently includes the most essential distinction between natural science and magic. It should be regrettably stated that forecasting as the final aim of the theories of natural science has partly escaping the attention of even the scientists themselves. It seems that this circumstance causes the passion felt sometimes for such diffuse formulations of those goals of scientific research which are sometimes noticeable as *explanation* or *revealing the essence* of phenomena.

As an example I can cite the caustically indicated (Kitaigorodsky 1978) tendency of chemists to *explain* a phenomenon with high precision by introducing after the event plenty adjusting parameters into formulas. A proper number of these can always achieve an ideal coincidence of the theoretical and the empirical curves, only not before the latter was experimentally obtained.

Kitaigorodsky (1978) offered a formula for quantitatively indicating the value P of a theory: $P = (k/n) - 1$. Here, k is the number of magnitudes which can be predicted by that theory, and n, the number of adjusting parameters. The value of a theory is therefore non-existent if $k = n$, and it is essential if k is much greater than n. The reader will be certainly justified to believe that this proposal is a joke, but of a kind that includes a large part of truth.

A somewhat exaggerated stress on the idea of forecasting noticeable in the newest discipline (*Prognostika* 1975/*Prognostication* 1978) is likely a reaction to the mentioned partial disregard of that fundamental idea. In this connection I indicate once more that in any concrete branch of natural science forecasting is not at all a novelty and that during many years a large and specific experience of forecasting had been acquired with a great deal of trouble. It is hardly possible to create some essentially new, general and at the same time substantial theory of forecasting. Meanwhile, however, a unification of terminology connected with forecasting can undoubtedly play some positive role.

2. The Initial Concepts of the Applied Theory of Probability

2.1. Random variables and their moments. Denote the controlled conditions of trials by U, their result by V and the magnitude measured in trial s by $X(s)$. The forecast of $X(s + 1)$ given $X(s)$ *often fails*. Permanence (forecast verified many times) is looked for by averaging and obtaining from initial unpredictable magnitude $V_1 = X(s)$

$$V_m = \mathrm{E}_m(X) = \bar{X}(s)$$

where $\mathrm{E}_m(X)$, in general also unpredictable, is the *empirical mean* of an unpredictable magnitude, of a random variable $X(s)$. It is often stable:

$$\mathrm{E}_m(X) \approx \mathrm{E}(X) \qquad (1)$$

which means that sooner or later the scatter of the values of $E_m(X)$ rather often appreciably diminishes. The author introduced *the pattern of an extended series of trials.* Bearing in mind his statements made in the sequel, it means that the behaviour of $E_m(X)$ is studied throughout the series rather than appreciated by the result of the last trial. This latter method is called *the pattern of a fixed series*. For a predictable permanence it is supposed that (1) persists when the series is extended and $E(X)$ is *the predictable* rough estimate of the empirical mean. Expectation of a separate measurement is meaningless.

The author introduces moments but barely applies them.

2.2. Statistical stability. It is often alleged that homogeneity of trials leads to statistical stability. Only *controlled* conditions of trials are meant and therefore, on the contrary, statistical stability means that the trials were homogeneous. Statistical stability is best justified by empirical induction. Without stability $E(X)$ does not exist.

Randomness (in the general sense) is identified with unpredictability. It became usual to understand random variables in the mathematical sense only as statistically stable unpredictable magnitudes, and even such for which the notion of distribution of probabilities is applicable.

This narrow specialized interpretation of random variable is still being willy-nilly confused with its wide general meaning and leads to a mistaken belief that the applied theory of probability and mathematical statistics are applicable to any random variable understood in the general sense, i. e., to unpredictable magnitudes.

On the other hand, the reader begins to believe that the mathematical propositions of the theory of probability somehow directly concern only such magnitudes. Actually, their unpredictability is not at all a necessary condition for applying to them the theory of probability. It is important that when measuring a magnitude many times it indicates statistical stability. An artificial introduction of unpredictability in an experiment by the so-called randomization as also in some calculations by the Monte Carlo method can be thought to mean an excessively brave challenge to the natural scientific tradition[1].

2.3. Probability of an event. An event A is random in both senses if $X_A(s)$ is random. Stability of frequency is established by empirical induction according to the pattern of an extended series. If frequency is stable, $E(X)$, the probability of an event, is its predictable rough estimate. If the behaviour of the series is not studied, and the probability only determined by its outcome, the statistical stability is not investigated.

Statistical probability is not applicable to individual trials. For estimating the probability of a rare event of the order of 10^{-4}, sometimes encountered in the reliability theory, 10^5 measurements are required.

2.4. Distribution of probabilities. It is measured for a series of an increasing number of trials. If the empirical distributions are stabilized, $F(X)$ is determined. This is empirical induction for the pattern of an extended series. Lack of stability of the empirical distributions means that the notion of $F(X)$ is not applicable. Often recommended is the measurement of those $F_m(X)$ because their stability is more noticeable, rather than the histograms, but this is akin to stating that an insensitive device is better than a sensitive histogram (indicating a greater scatter, a lack of stability).

2.5. Statistical independence. Lack of correlation. A necessary and sufficient condition of independence is

$$F(X_1, X_2, ..., X_n) = F_1(X_1)\, F_2(X_2) \ldots F_n(X_n).$$

It does not exist always even if the pertinent magnitudes are intuitively independent. Statistical independence can only be discussed with complete justification after establishing statistical stability. Independence of a separate measurement is meaningless. Non-correlation of pairs of magnitudes $X_1(s)$, ..., $X_n(s)$ means that $E(X_i, X_j) = 0$ for $i, j = 1, ..., n$ and $i \neq j$.

2.6. The main problem of the applied theory of probability. After heuristically forecasting the initial magnitude V, to predict theoretically some secondary magnitude, their functions. Forecasting the initial magnitudes is always intuitive.

2.7. Limit theorems of the theory of probability. For the central limit theorem (CLT) magnitudes $X_1(s)$, ..., $X_n(s)$ are considered statistically independent for any n and their scatter around their expectations is supposed to be roughly the same. For the law of large numbers (LLN) the second demand is dropped and the first one weakened so that variance can be even replaced by non-correlation. The CLT is practically admitted if the LLN takes place intuitively. Quantitative estimates during the proof of the laws of large numbers are only possible by means of the [Bienaymé –] Chebyshev inequality but they are rough and inexpedient as compared with the CLT. In the initial period of the development of the probability theory the fundamental importance of limit theorems had been essentially exaggerated which is not completely done away with even now[2]. Thus, sometimes statements are made asserting that statistical stability is due to the LLN.

2.8. The Mises approach. His initial concepts are extremely close to being experimental. Instead of stability of the empirical mean he postulates the existence of

$$E(X) = \lim E_m(X), \; m \to \infty.$$

The pattern of an extended series is meant here. Particular cases are the definition of probability as the limit of frequency and of $F(X)$ being the limit of $F_m(X)$. The convergence can be understood in different ways.

Randomness (that is, unpredictability) does not enter directly, the whole arsenal of tools is typically mathematical. In similar ways, mathematicians discuss derivatives and integrals rather than velocities or specific heat. Transitions to the limit are only the means (or necessary! expenses) of a rigorous formalization[3]. The Mises approach provides civil rights in the theory of probability for the known empirical patterns of treating data dating back to the very foundations of the natural scientific method with its demand of repeated reproduction of results.

The main feature of the Mises approach consists in dealing with everything as though considering an experiment. Not surprisingly, expectation is introduced in applications according to his postulate often without citing Mises.

2.9. Comparison with the Kolmogorov axiomatization. The Mises approach most likely can not be included within the boundaries of this axiomatization. The main theoretical problem apparently consists in discovering existence theorems for number sequences converging to the given beforehand distribution function. This problem is still only solved for weak convergence (Postnikov 1960). The Mises approach should be specially developed by number-theoretic methods unusual for the Kolmogorov axiomatics.

The foundations of the Mises approach can be quite rigorously formed as a clear set of axioms. Contrasting it to the axiomatic method is wholly based on a misunderstanding.

2.10. Conclusion. Unpredictability of repeatedly measured initial magnitudes is neither necessary, nor sufficient for enlisting the theory of probability since it does not ensure the initial statistical stability, i. e. the stability of the averaged characteristics of the initial magnitudes, which is the really necessary pertinent condition. Independence of trials is often presumed, but see § 2.5.

I did not have an occasion to enlist officially the notion of independence of trials. The introduction of controlled conditions U into quantitative notions, formulas or propositions of the theory of probability however constructed apparently can not be even hoped for. The introduction of the concept of independence of trials is not required by the notions of statistical stability and statistical independence of magnitudes. On the contrary, it should be based on these notions.

3. Critical Analysis of the Method of Mathematical Statistics

According to one of the usual definitions (Nikitina et al 1972), *mathematical statistics studies quantitative relations of mass phenomena* [...] *It is closely linked with the theory of probability.* [...] *Its methods are universal*[4].

3.1. An alternative to the general purpose of mathematical statistics. All treatises state that that purpose is to provide a universal numerical theory of measuring averaged characteristics; to find out whether a given sample is representative.

The possibility of constructing such a theory is doubtful since the precision and reliability of the *initial* presumptions can hardly be calculated or justified. Those presumptions are *intuitive forecasts* of some permanences. When adopting them, the alternative is to abstain as much as possible from theoretical considerations, to substantiate their likelihood of forecasts by empirical induction. We will discuss how to verify experimentally the typical pronouncements made by mathematical statistics.

3.2. Traditional interpretation of limit theorems. [Only the Bernoulli theorem is discussed.] It only deals with *one* series of observations and applies two fundamental notions of mathematical statistics, independence of trials and convergence in probability.

3.3. Independence of trials. Contrary to what is sometimes asserted, stability of the controlled conditions of the experiments is not sufficient and the conditions U can not at all quantitatively enter the theory of probability. It is less superficially stated that each trial engenders a random variable so that the independence of the n trials is reduced to the statistical independence of the n variables.

However, a trial (a measurement of X) only engenders a realization of a random variable, the number $X(s)$. Mathematical statistics has no clear rules for empirically verifying independence of trials, for discussing an ensemble of such series. The correspondence n trials – n random variables means imagining an ensemble of random variables.

Such imagining is a peculiar feature of mathematical statistics, and there are no clear rules for empirically verifying the results of the trials.

3.4. Convergence in probability. For experimentally checking it[5] a long series of secondary trials is required and many samples of size n are needed. The author calls forming many samples the *pattern of many series*, and the patterns of an extended and a fixed series are now both called the *pattern of one* (extended or fixed) *series*. In mathematical statistics, an ensemble of sequences of trials is only imagined.

3.5. Two competing mathematical models of statistical stability. Thus, the traditional formulation of the limit theorems lack clear rules for verifying either the conditions, or conclusions. This is the reason that had formerly engendered an illusion, not completely dissociated from, that the laws of large numbers theoretically deduce stability of means from homogeneity of trials. In particular, it followed that mathematical statistics identifies statistical stability with convergence in probability as studied in the laws of large numbers. The Mises model of stability $P = \lim \omega$, $n \to \infty$, is not usually mentioned. The author quotes Kolmogorov's pertinent remark (1956, p. 262):

Such considerations can be repeated an unrestricted number of times, but it is quite understandable that it will not completely free us from the necessity of turning during the last stage to probabilities in the primitive, rough understanding of that term[6].

To put it otherwise, there is no other way out except turning to the pattern of one series, i. e. to the Mises model of stability of frequencies. If you wish, the Mises definition of probability is exactly the turn to probabilities *in the primitive rough understanding of that term*. According to common sense, the turn to the last stage should be done in such a manner that the probabilities of the highest rank included in the mathematical model of the given experiments were indeed actually measured in that experiment. It is apparently difficult to warrant the imagination of probabilities of even one superfluous rank. Nevertheless, such imagination is one of the fundamentals of the method of mathematical statistics.

3.6.1. Postulate of the existence of a distribution of probabilities for the initial random variables. All the considerations in mathematical statistics usually begin by postulating the existence, and sometimes even the concrete type of the distribution of probabilities for unpredictable magnitudes, then the *estimation* of density or parameters of the *objectively existing* distribution is demanded. The Fisherian theory of estimation is constructed according to this pattern as also the method of maximal likelihood, the theories of confidence intervals, of order statistics etc[7]. An alternative (see Chapter 2) is to concentrate on empirical justification of predictions of statistical stability.

The most difficult and interesting problem of empirically investigating statistical stability is rapidly sped by. Here is Grekova's critical remark (1976, p. 111) about calculating a confidence interval when the number of trials is small:

A rather subtle arsenal is developed based on the assumption that we know the distribution of probabilities of the random variable (the normal law). And once more the question emerges: wherefrom indeed do we know it? And how precisely? And, finally, what is the practical value of the product itself, of the confidence interval? A small number of trials means small amount of information, and things are bad for us. But, whether the confidence interval will be somewhat longer or shorter, is not so important the less so since the confidence probability was assigned arbitrarily.

From my viewpoint, this remark is still a rather mild doubt. We may add: Wherefrom and how precisely do we know that, given this concrete situation, it is proper at all to discuss distributions of probabilities? Suppose, however, that the distribution of probabilities of the unpredictable magnitudes under discussion does exist. But then (Grekova 1976), it is not necessary to think highly of the theory of estimation. Indeed, this theory allows us to extract the maximal amount of information not from sample data in general; the postulate on the type of distribution of probabilities is also introduced. It only represents reality with some precision at whose empirical estimation the estimation theory is not at all aimed.

And the theory's conclusions and it itself, generally speaking, changes with the change of that distribution. It would have been necessary to calculate the vagueness of the sought estimates of the parameters caused by the expected vagueness of the postulated distribution. Then, the estimation theory extracts the maximal amount of information according to some specific criteria whose practical value is not doubtless. Finally, that theory is based on the postulate of independent trials with which, as we saw, not everything was in order. It ought to be stated that the treatises on mathematical statistics do not miss the opportunity to identify the treatment of observations, that really not at all simple discipline, with the scientific approach in statistics. Here is Grekova (1976, p. 112) once more:

Mathematical arsenals have some hypnotic property and researchers are often apt to believe unquestionably their calculations, and the more so the more flowery are their tools [...].

In any applied science, a scientific approach presumes first of all a creation of an intuitively convincing empirical foundation. The complication, rigour and cost of the mathematical arsenal should be coordinated with the reliability of the foundation. This pragmatic rule applied from long ago is neatly called *principle of equal stability of all the elements of an* [applied – Yu. A.] *investigation* (Grekova 1976, p. 111). The theory of estimation hardly satisfies it in due measure.

3.6.2. Postulate on the existence of a distribution of probabilities for sample estimates. Imagining many additional samples. The existence and sometimes even the type of that distribution is postulated. Suppose that an experiment according to the pattern of many series is carried out. We may only repeat what was said in § 3.6.1 concerning the distributions of the initial random variables.

Actually, however, only the parameter of the distribution is studied. Its estimate is usually found by treating *all* the data as a single entity. In mathematical statistics, this procedure is accompanied by *imagining* many additional samples, presuming the postulate of § 3.6.1 and independence of the trials.

The alternative is to discuss, as far as possible, only random variables really measured in *long* series of trials and to keep to the pattern of one extended series. When several series are available, the method of maximal likelihood will provide several optimal estimates, so which is the most optimal? Not less strange will be the concept of confidence interval.

3.6.3. Postulate on independence of trials. For mathematical statistics, it occupies in some sense a central position because it links the postulates of §§ 3.6.1 and 3.6.2. However, it is hardly elementary, see Chapter 4.

3.7. The choice of a threshold for discerning. In its very essence it is intuitive and unavoidable for verifying and comparing various statistical hypotheses with each other. Mathematical statistics can not naturally avoid it, but only shifts the choice to magnitudes not being measured in reality. No special benefit is seen in that procedure.

3.8. The problem of representativeness of samples. To all appearances, this should be frankly attributed to a problem non-formal in its very essence, to the choice of the initial intuitive assumptions. An alternative can be to separate the trials into *several* subsamples and only forecast rough averaged characteristics. The size of the subsamples and the threshold for discerning should be chosen according to precedents in a candid intuitive way in terms of measured magnitudes. Such an empirical intuitive approach embodies the fundamental principle of natural science, the demand of *multiple* repetition of experiments and a convincing reproduction of their results. See Alimov (1976, 1977, 1978b; 1978a; 1979).

4. The Mises Formalizations of the Idea of Independent Trials

In § 3.3 we concluded that a clear rule is required for transition from one initial sequence of trials to an ensemble of statistically independent sequences. That rule should somehow reflect intuitive ideas about independence of trials. We may accept Mises' general idea to consider the trials independent if their sequence is very irregular and difficult to forecast. He called such sequences *irregular collectives*.

From the 1920s many authors (Wald, Feller, Church, Reichenbach) had developed various versions of formalizing the concept of such collectives. Kolmogorov's algorithmic notion of probability of 1963[8] also bears relation to this problem although it is apparently only indirectly linked with the idea of forecasting. See the pertinent initial bibliography, for example, in Knut (1977, vol. 2, chapter 3).

4.1. Formalization according to Ville [e. g., Shafer & Vovk 2001, pp. 48 – 50] **and Postnikov (1960).**

4.2. Formalization according to Copeland. Postnikov (1960) proved that a sequence is irregular in Copeland's sense if and only if it is irregular according to Ville and Postnikov.

4.3. General remarks on §§ 4.1 and 4.2. A sequence irregular according to §§ 4.1 or 4.2 presents a simplest example of an intuitive and rigorous mathematical model of trials which can be called independent and identical (identical since the distributions of the probabilities for all the formed sequences coincide). The idea of a poor predictability of one initial sequence is here indeed reduced rather naturally to demanding statistical independence of the ensemble of sequences. As a result, independence of trials is treated in such a manner that provides a sufficiently clear rule for its quantitative empirical verification.

Thus, after being clearly formulated, independence of trials obviously becomes a concept derived from the notion of statistical stability, cf. our assumption in § 2.10. It follows that the postulate of § 3.6.3 even in its most simple clear form is evidently more complex than the postulates of §§ 3.6.1 and 3.6.2. It can not be the assumption from which, at least according to the pattern of one series, statistical stability is deduced.

The verification of any propositions of mathematical statistics will be therefore aimed at verifying the postulate of § 3.6.3 rather than at measuring the sought parameters of the distributions of the initial magnitudes. This measurement, for which, as it seems, mathematical statistics is indeed created, will only constitute a small and so to say preliminary part of the work to be done.

The formulations of the idea of independence of trials considered above are obviously only applicable when the *n* trials are actually carried out *many* times. The alternative to the method of mathematical statistics therefore means that the postulate of § 3.6.3 should be introduced only after the sought parameters or the initial distribution itself were reliably measured.

4.4. Specification of the traditional formulations of the limit theorems on the basis of the concept of an irregular collective. The author interprets the Bernoulli theorem by applying the notion of irregularity of collectives. One of the conditions of his pertinent theorem is the existence of a limit of the sequence of trials, the probability according to Mises.

He notes that his (and therefore the Bernoulli) theorem does not claim to justify the statistical stability of the frequency which is now one of his preconditions. He concludes that the limit theorems (in general!) are not actually *fundamental propositions* as it was thought in the initial period of the development of the theory of probability.

4.5. An example from classical statistical physics. [Concerning the work of an oscillator being in thermal equilibrium with a thermostat.]

5. Conclusion

An alternative to the method of mathematical statistics can be described in a few words in the following way. In applied research, and more precisely beyond fundamental physics, we should as far as possible abstain from introducing stochastic magnitudes not measured in real experiments in our initial assumptions. The so-called numerical experiments compare a computer and a *paper* model but not model and reality.

The objects of study in economics, sociology and even modern technology are most often too complicated and unstable for constructing their useful models by issuing from general principles

peculiar for the foundations of physics but remote from experiment. Advisable here are efficient phenomenal models without special claims to fundamentalism. According to the principle of equal stability of all the elements of an applied investigation, introduction of complicated mathematics should be considered guardedly. I conclude by quoting Wiener (1966 from Russian), hardly an opponent of mathematization:

Advancement of mathematical physics caused sociologists to be jealous of the power of its methods but was hardly accompanied by their distinct understanding of the intellectual sources of that power. [...] *Some backward nations borrowed Western clothes and parliamentary forms lacking personality and national distinctive marks, vaguely believing as though these magic garments and ceremonies will at once bring them nearer to modern culture and technology, − so also economists began to dress their very inexact ideas in rigorous formulas of integral and differential calculuses.* [...] *However difficult is the selection of reliable data in physics, it is much more difficult to collect vast economic or sociological information consisting of numerous series of homogeneous data.* [...] *Under these circumstances, it is hopeless to secure too precise definitions of magnitudes brought into play. To attribute to such magnitudes, indeterminate in their very essence, some special precision is useless. Whatever is the excuse, application of precise formulas to these too freely determined magnitudes is nothing but a deception, a vain waste of time*.

Notes

1. Both randomization and the Monte Carlo method are mentioned by Prokhorov (1999) and Dodge (2003). Tutubalin, who had sided with Alimov, later applied the Monte Carlo method in a joint contribution (Tutubalin et al 2009, p. 189). O. S.

2. Concerning the theory of probability the author was likely wrong, see Tutubalin [i, § 4.2], who [i, § 4.5] also remarked that for natural science the significance of the LLN only consisted in reflecting the experimental fact of the stability of the mean. The author's next sentence had to do with the application of the LLN to statistics, but he only stated what that theorem did not achieve.

Concerning the CLT I quote Kolmogorov (1956, p. 269): *Even now, it is difficult to overestimate* [its] *importance*. O. S.

3. In spite of numerous efforts made, the Mises approach remains actually questionable, see end of [vi]. O. S.

4. It is worthwhile to quote another definition (Kolmogorov & Prokhorov 1974/1977, p. 721):

[Mathematical statistics is] *the branch of mathematics devoted to the mathematical methods for the systematization, analysis and use of statistical data for the drawing of scientific and practical inferences.* O. S.

5. See the Introduction to [v]. O. S.

6. I illustrate principal and secondary magnitudes (§ 2.6) by Kolmogorov's reasoning. Frequency μ/n tends to probability p, and the probability $P(|\mu/n - p| < \varepsilon)$ is a secondary magnitude which in turn should be measured as well. O. S.

7. This statement is not altogether correct. See Wilks (1962, Chapter 11) and Walsh (1962) who discuss non-parametric estimation and order statistics respectively. O. S.

8. Perhaps Kolmogorov (1963). O. S.

Bibliography

Alimov Yu. I. (1976, 1977, 1978b), *Elementy Teorii Eksperimenta* (Elements of the Theory of Experiments), pts 1 – 3. Sverdlovsk.

--- (1978a Russian), On the applications of the theory of probability considered in V. N. Tutubalin's works. *Avtomatika*, No. 1, pp. 71 – 82.

--- (1979 Russian), Once more about realism and fantasy in the applications of the theory of probability. Ibidem, No 4, pp. 103 – 110.

Anscombe F. J. (1967), Topics in the investigation of linear relations. *J. Roy. Stat. Soc.*, vol. B 29, pp. 1 – 52.

Blekhman I. I., Myshkis A. D., Panovko Ya. G. (1976), *Prikladnaia Matematika* (Applied Math.). Kiev.

Dodge Y. (2003), *Oxford Dictionary of Statistical Terms*. Oxford.

Grekova I. (1976 Russian), Special methodical features of applied mathematics on the current stage of its development. *Voprosy Filosofii* No. 6, pp. 104 – 114.

Kitaigorodsky A. I. (1978), *Molekuliarnye Sily* (Molecular Forces). Moscow.

Knut D. E. (1977 Russian), *The Art of Computer Progamming*, vol. 2. Moscow. The author referred to this Russian edition.

Kolmogorov A. N. (1956 Russian), Theory of probability. In: *Matematika. Ee Soderzanie, Metody i Znachenie* (Mathematics. Its Contents, Methods and Importance), vol. 2. Moscow, pp. 252 – 284.

Kolmogorov A. N., Prokhorov Yu. V. (1974 Russian), Statistics. *Great Sov. Enc.*, 3rd ed., English version, 1977, vol. 15, pp. 721 – 725.

Nikitina E. P., Freidlina V. D., Yarkho A. V. (1972), *Kollekzia Opredeleniy Termina "Statistika"* (Collection of Definitions of the Term "Statistics"). Moscow.

Postnikov A. G. (1960), *Arifmeticheskoe Modelirovanie Sluchainykh Prozessov* (Arithmetical Modelling of Stochastic Processes). Moscow. Perhaps included in author's *Izbrannye Trudy* (Sel. Works). Moscow, 2005.

Prognostication (1975 Russian), *Great Sov. Enc.*, 3rd ed., English version, vol. 21, 1978.

Prokhorov Yu. V., Editor (1999), *Veroiatnost i Matematicheskaia Statistika. Enziklopedia* (Probability and Math. Statistics. An Enc.). Moscow.

Shafer G., Vovk V. (2001), *Probability and Finance. It's Only a Game*. New York.

Tutubalin V. N. (1972), *Teoria veroiatnostei* (Theory of Probability). Moscow.

Tutubalin V. N., Barabasheva Yu. M., Devyatkova G. N., Uger E. G. (2009 Russian), Kolmogorov's criteria and Mendel's heredity laws. *Istoriko-Matematich. Issledovania*, ser. 2, No. 13/48, pp. 185 – 197.

Venikov V. A. (1978), *Perekhodnye Elektromekhanicheskie Prozessy v Elektricheskikh Sistemakh*. Moscow.

Vysotsky M. (1979), *Pod Znakom Integrala* (Under the Sign of Integral). Moscow.

Walsh J. E. (1962), Nonparametric confidence intervals and tolerance regions. In: Sarhan A. E., Greenberg B. G., Editors, *Contributions to Order Statistics*. New York – London, pp. 136 – 143.

Wiener N. (1966 Russian), *Tvorez i Robot* (Creator and Robot). Moscow. The author referred to this Russian edition. German: possibly *Mensch und Menschenmaschine*.

Wilks S. S. (1962), *Mathematical Statistics*. New York.

V

V. N. Tutubalin

Answering Alimov's Critical Comments on Applying the Theory of Probability

Otvet na kriticheskie zamechania Yu. I. Alimova
v sviazi s problemami prilozenia teorii veroiatnostei.
Avtomatika, No. 5, vol. 8, 1978, pp. 88 – 91

Introduction by the Translator: The Main Ideas of Alimov (1978)

Page 71. The stability of the initial means is a postulate whose likelihood should be experimentally justified.

Page 73. The LLN was, and sometimes still is considered a bridge connecting the theory of probability with practice. According to the context (p. 72), the author denies this statement because statistical stability of the trials had to be proved.

Page 73. The proximity of the empirical frequency to the initial probability should be estimated by measurement.

Page 74. Not practice is following Mises as Tutubalin remarked, but rather the inverse had happened.

Page 75. The significance of the LLN and other limit theorems in statistics is reduced to solving an ordinary problem.

Page 76. An explanation of the independence of trials is not fundamentally important for the Mises approach. Statistical independence can be revealed by most various sequences of trials including periodic sequences.

Page 77. For applications, the transition of the empirical frequency to probability is an undistinguished expense of a rigorous formalization rather than any essential feature of the Mises approach.

Page 77. Without due substantiation but in agreement with the former pronouncement the author alleges that *the so-called strong laws of large numbers* are very remote from the theory of probability.

[The main text]

Alimov (1978) critically commented on some of my publications and his paper is the only one that I know to publish a response to my methodical and popular scientific works. Since discussions, including those carried out in public, are most necessary for the development of science and teaching, the initiative of the periodical *Avtomatika* as well as the serious (as will be seen below) work of Alimov only due to which that discussion became possible should be appreciated very positively.

Alimov is well known because of a number of his publications, mostly of a critical kind, on the application of the probability theory. I think that the general aim of his contributions differ but little from mine. We both apparently agree that the amount of falsehoods arrived at by *applying* the theory of probability is too great to be tolerated. In a historical perspective, my statement made publicly is all by itself a quite effective means of combating that evil. And indeed intrinsic

processes are now going on in the society due to which the part played by moral elements sharply increases. It is this circumstance to which I and Alimov are beholden for some not excessive popularity of our publications; otherwise they would just have not been popular.

Thus, connecting the problem of the truth of scientific work in the first place with the level of social morals, I consider the possibility of solving that problem by purely scientific means rather sceptically, for example by describing the theory of probability according to Mises rather than by the generally recognized Kolmogorov axiomatics. I do not mention the idea of official censuring voiced by Alimov (1978, p. 82). That would have been only really helpful if those responsible will be at the same time as though automatically endowed with the truth or at least with a tendency to it.

Incidentally, I would like to turn Alimov's attention to a circumstance which I myself previously experienced, that apparently any attempt to retell or cite the viewpoint of other people introduces unavoidable corruption. Thus, Alimov (1978, p. 74) says: *When comparing the Mises approach with a dead language, Tutubalin nevertheless notes ...*
Actually, I (1972, p. 148) wrote:

> *In general, the present attitude of specialists towards the language of the Mises theory can be compared with the attitude towards a dead language in which for some reason no one wishes to speak although, after being appropriately corrected and altered, it will be quite capable of expressing everything spoken in a live language*.

Thus, after Alimov quite properly but [too] briefly arranged my viewpoint, my friendly attitude towards probability theory according to Mises absolutely disappeared and became replaced by disdain.

This example taken together with my general opinion about the corruptions of such a kind being practically unavoidable, sufficiently explains why I do not reply in detail to each point of Alimov's criticism. Concerning general pronouncements, all is reduced to selecting some shade of conception. For example, if Alimov [p. 75] thinks that the law of large numbers is a limit theorem suited for solving an ordinary modest problem of probability theory unconnected with the principles of its applicability, then let him be in the right.

However, it is much more interesting to turn to concrete examples of application of the theory because they are always richer. For example, prominent physicists who had been creating that science usually philosophically interpreted it themselves without needing philosophers. Not that I deny the social utility of philosophers, but their customers are not leading scientists but the multitude of those who do not (yet) occupy leading places in science.

Alimov's main merit as a critic, as it seems to me, is that he considered concrete numerical data. I bear in mind the experimental verification of the most simple Mendelian law of assortment of indications in the ratio 3:1. The data was provided by Ermolaeva (1939), a representative of the Lyssenko school, and Enin (1939), its opponent. Kolmogorov (1940) published a detailed analysis of

Ermolaeva's results and concluded that, instead of refuting the Mendelian law, she completely confirmed it. There also, without minutely analysing Enin's paper, Kolmogorov implied that his results are doubtful because they confirmed that law too finely.

In a popular scientific booklet, I [iii] thought it expedient to remind readers about Kolmogorov's paper and supplemented it by treating Enin's results. Alimov treated the same data otherwise and formulated a number of objections. He directed them to me alone although a part of them to the same extent concerned Kolmogorov's calculations. I begin with the objection which I understand and consider essential.

He notes that in many cases the *families* considered by Ermolaeva were small (not more than 10 observations). Then the normal approximation of the frequencies of a certain phenotype introduced by Kolmogorov ought to be very rough. In particular, the presence of normed frequencies smaller than − 3 which I [iii] considered as significant deviations from the Mendelian law can be explained. as Alimov believes, by the asymmetry of the binomial law. Alimov declared that my conclusion was wrong (that was somewhat hastily, he should have said *unjustified*). *Any student of a technical institute*, as he states, would have avoided such a mistake caused by the general corruption of concepts due to the application of the *non-Mises language and the rituals of mathematical statistics*.

Actually, everything is much simpler. Before preparing my booklet, I did not acquaint myself with Ermolaeva's paper which was not readily available. Now, however, since her data became an object of discussion, I had a look at that source. The data on the assortment in separate families are provided there in Tables 4 and 6. In Table 4 the families are numbered from 1 to 100, but for some unknown reason numbers 50 and 87 are omitted. In Table 6, the numbering begins with 22 and continues until 148, but numbers 92, 95, 115, 127, 144 are absent. At the same time, the table showing the total, states 100 and 127 families respectively.

Kolmogorov inserted a venomous pertinent remark; he counted 98 families in the first, and 123 (actually, 122) in the second table. The general style of her contribution, let me say it frankly, is abominable. The author obviously does not understand the meaning of the errors calculated by *biometric methods* for the number of assortments. It is quite clear that her data do not really deserve to be seriously considered.

However, if only discussing Ermolaeva's tables such that they are, Alimov is still unjustly reproaching me for discovering non-existent deviations from the Mendelian law. Indeed, Table 4 includes a result of assortment 0:17, and 0:10 in Table 6 instead of the expected ratio 3:1. Their probabilities are 4^{-17} and 4^{-10} respectively so that, having 200 plus trials, such events could not have occurred.

Concerning both Kolmogorov's and my own treatment, I would like to indicate that, in spite of Alimov' opinion, correct scientific results are possibly often obtained not because we do everything properly, but owing to some special luck.

I did not understand the meaning of Alimov's objection to the calculation of the confidence level. From the times of Laplace, after

obtaining a deviation from the theory assumed to be valid, scientists have been attempting to calculate, if possible, the probability of a deviation not less than that. If that probability was high, 1/2, say, everything was in order; otherwise, supposing that its order was 1/1000, it was advisable to look for the cause of the deviation. If, finally, it was moderate, its order being 1/10, say, the case was doubtful and a final decision impossible. Can we object to such kind of applying the confidence level?

I do not understand Alimov's concept of independence either. On p. 80 he thinks that *secondary* trials, that is, data on the assortment of indications in different families, unconnected with each other, can be statistically dependent. But how could that occur with the outcomes of different trials unconnected with each other? If as a result of *one* trial events A and B can either happen or fail, they can be statistically dependent and, when treating this dependence according to Mises, we should use a single record. But in case of two absolutely different trials we should apparently introduce something like a direct product of two records.

Finally, concerning my treatment of Enin's data, Alimov remarks first of all that his number of families is so small (11 + 14 = 25), that their treatment did not warrant the waste of either time or paper with a special non-linear scale. I will answer that by stating that, on the contrary, I aimed at showing that the image of a distribution function unlike that of a histogram allows to obtain sensible results even when having such a small sample size.

Then, Alimov states that it was possible to arrive at my conclusions by compiling an extended sample[1]. To some extent this is correct, but to some extent wrong. After taking samples of about the same size, the frequencies in Enin's second sample will be closer to the theoretical magnitudes than Ermolaeva's similar frequencies. This is seen in Alimov's table (1978, p. 78). It can be therefore concluded, if Ermolaeva's data are considered as a standard, that there is some trouble with Enin's materials.

However, after calculating the chi-squared statistic (Tutubalin [iii]), a standard is not needed. Actually, Alimov (1978, pp. 80 – 81) believes that Enin's data should be treated not by means of the normal distribution of the normed frequencies, but by a more subtle model. In principle, I completely agree, only that model should not be a mixture of binomial distributions (Alimov, p. 80, formula (21)), but it should directly consider the actual numerical strength of the families. A series of binomial trials would be obtained having a known number of trials and a known probability of success. Understandably, such a model is barely convenient and therefore the stupidest Monte Carlo method[2] will apparently be most effective for calculating the various pertinent probabilities. Thus, for example, the true distribution of the Kolmogorov statistic or some other statistic measuring the deviation from the Mendelian law can be determined. Since such statistics are rather diverse, we conclude that not only the electron or the atom but also the certainly carelessly constructed Ermolaeva's tables are inexhaustible[3].

Notes

1. Alimov [iv, § 2.1] introduced *extended* series of observations. O. S.

2. Without saying anything else, I note that Tutubalin himself applied that method in a joint paper (Tutubalin et al 2009, p. 189). O. S.

3. That the electron is inexhaustible is Lenin's celebrated statement from his *Materialism and Empirical Criticism* (1909, in Russian). The notion of electron is intrinsically contradictory, so perhaps the author indirectly stated the same about those tables. Anyway, Lenin's statement remains unjustified. O. S.

Bibliography

Alimov Yu. I. (1978 Russian), On the problem of applying the theory of probability considered by V. N. Tutubalin. *Avtomatika*, No. 1, pp. 71 – 82.

Enin T. K. (1939 Russian), The results of an analysis of the assortment of hybrids of tomatoes. *Doklady Akademii Nauk SSSR*, vol. 24, No. 2, pp. 176 – 178. Also published at about the same time in a foreign language in *C. r. (Doklady) Acad. Sci. URSS*.

Ermolaeva N. I. (1939 Russian), Once more about the "pea laws". *Jarovizatsia*, No. 2, pp. 79 – 86.

Kolmogorov A. N. (1940), On a new confirmation of Mendel's laws. *C. r. (Doklady) Acad. Sci. URSS*, vol. 28, No. 9, pp. 834 – 838.

Tutubalin V. N. (1972), *Teoria Veroiatnostei* (Theory of probability). Moscow.

Tutubalin V. N, Barabasheva Yu. M., Devyatkova G. N., Uger E. G. (2009 Russian), Kolmogorov's criteria and verification of Mendel's heredity laws. *Istoriko-Matematich. Issledovania*, ser. 2, issue 13/48, pp. 185 – 197.

VI

Oscar Sheynin

On the Bernoulli Law of Large Numbers

Bernoulli considered (independent) trials with a constant probability of *success*, and rigorously proved that the frequency of success tends to that probability. Mises, however, treated collectives, totalities of phenomena or events differing from each other in some indication, and characterized by the existence of the limiting frequency of success and by irregularity. The latter property meant that for any part of the collective that limiting frequency was the same.

Alimov noted that artificially constructed collectives proved that the empirical frequency of success can become more stable as the number of trials increased, but have no limit. Therefore, the existence of that limit is an experimental fact. I have described his viewpoint in some detail in an Introduction to [v]. Tutubalin largely sided with Alimov.

In the same *Ars Conjectandi*, previous to proving the LLN, Bernoulli stated that his law was also valid in its inverse sense (and De Moivre independently stated the same with respect to the first version of the CLT proved by him in 1733). In other words, an unknown and even a non-existing probability (one of Bernoulli's examples) could be estimated by the limiting frequency.

In a little known companion paper (1765) to his main memoir (1764), Bayes all but proved his own limit theorem explicating that inverse LLN. He did not make the final step from the case of a large finite number of trials because he opposed the application of divergent series which was usual in those times. That was done in 1908 by Timerding, the Editor of the German translation of Bayes, certainly without using divergent series.

Bayes – Timerding examined the behaviour of the centred and normed random variable η, the unknown probability, $(\eta - E\eta)/\mathrm{var}\,\eta$ whereas the direct LLN dealt with the frequency ξ, $(\xi - E\xi)/\mathrm{var}\xi$. His main memoir became widely known and for a long time the Bayes approach had been fiercely opposed, partly because an unknown constant was treated as a random variable (with a uniform distribution). Note that $\mathrm{var}\eta > \mathrm{var}\xi$ which is quite natural since probability is only unknown in the inverse case. For attaining the same precision the inverse case therefore demands more trials than the direct law. Mises could have called Bayes his main predecessor; actually, however, he only described the work of the English mathematician, and inadequately at that. Bayes completed the first stage of the development of probability theory.

Alimov's viewpoint was largely correct since he considered an incomparably more general pattern than Bernoulli and thought about the necessary checks, but he [iv] was too radical in denying important parts of mathematical statistics as also too brave in altering the Mises approach. To borrow an expression from Tutubalin [end of ii], he introduced the Mises approach *of a light-weighted type.*

Concerning the rigor of the frequentist theory, witness Uspensky et al (1990, § 1.3.4):

Until now, it proved impossible to embody Mises' intention in a definition of randomness that was satisfactory from any point of view.

I ought to add, however, that Kolmogorov (1963, p. 369) had essentially softened his viewpoint about that theory:

I have come to realize that the concept of random distribution of a property in a large finite population can have a strict formal mathematical exposition.

In the 19th and 20th centuries statisticians had been reluctant to justify their studies by the Bernoulli LLN. They did not refer either to the inverse law or to Poisson (which would not have changed much). Maciejewski (1911, p. 96) even introduced *la loi des grands nombres des statisticiens* that only stated that the fluctuation of statistical numbers diminished with the increase in the number of trials. Romanovsky (1924, pt 1, p. 15) stressed the natural scientific essence of the LLN and called it physical. Chuprov (1924, p. 465) declared that the LLN included either mathematical formulas or empirical relations and in his letters of that time he effectively denied that the LLN provided a bridge between probability and statistics.

Bibliography

Bayes T. (1764), An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc*., vol 53 for 1763, pp. 360 – 418. Reprint: *Biometrika*, vol. 45, 1958, pp. 293 – 345.

--- (1765), Demonstration of the second rule in the essay [of 1764]. Ibidem, vol. 54 for 1764, pp. 296 – 325.

Chuprov A. A. (1924), Ziele und Wege der stochastischen Grundlagen der statistische Theorie. *Nord. Stat. Tidskr.*, t. 3, pp. 433 – 493.

Kolmogorov A. N. (1963), On tables of random numbers. *Sankhya, Indian J. Stat.*, vol. A25, pp. 369 – 376.

Maciejewski C. (1911), *Nouveaux fondements de la théorie de la statistique.* Paris.

Romanovsky V. I. (1924 Russian), Theory of probability and statistics according to some newest Western scholars. *Vestnik Statistiki*, No. 4 – 6, pp. 1 – 38; No. 7 – 9, pp. 5 – 34.

Uspensky V. A., Semenov A. L., Shen A. Kh. (1990 Russian), Can an (individual) sequence of zeros and ones be random? *Uspekhi Matematich. Nauk*, vol. 45, pp. 105 – 162. This periodical is being translated cover to cover.