Privately printed as a manuscript

B. V. Gnedenko, A. Ya. Khinchin

An Elementary Introduction to the Theory of Probability

Б. В. Гнеденко, А. Я. Хинчин

Элементарное введение
в теорию вероятностей

Большое число изданий начиная с 1946 г.

Translated by Oscar Sheynin

Berlin

2015

Contents

Foreword to the Fifth Edition

I have prepared this edition after the death of Khinchin, an eminent scientist and pedagogue. His name is connected with many ideas and results of modern probability theory. To him is due a systematic application of the methods of the set theory and the theory of functions of a real variable in probability theory; the construction of the fundamentals of the theory of random processes; an extensive development of the theory of summation of independent random variables; and the development of a new approach to the problems of statistical physics and of a harmonious system of its exposition. Together with S. N. Bernstein and A. N. Kolmogorov, Khinchin shares the merit of constructing the Soviet school of probability theory which is playing an outstanding role in modern science.

I am happy to have been his student. We wrote this book when the Great Patriotic War had been victoriously ending and our examples reflected elementary military problems. Now, fifteen years after our victory, when the country is covered by scaffolds of new buildings, it is natural to extend the scope of those examples. It is exactly for this reason that, without changing the general exposition or the elementary essence of the book, I allowed myself to substitute new examples for many of the previous ones. With a few exceptions I made the same alterations in the French edition of this book (Paris, 1960).

Moscow, 6 October 1960. B. V. Gnedenko

Foreword to the American Edition

In recent years, the theory of probability has acquired exceptionally great importance for the development of mathematics itself as well as for the progress of literally all branches of natural science, technology and economy. Its role is now beginning to be acknowledged in linguistics and even in archaeology. It is for this reason that it is essential to popularize its ideas and results as widely as possible and in all their varieties.

In many countries there is a persistent demand for the introduction of the elements of the theory of probability into the high-school curriculum. This point of view was also shared by A. Ya. Khinchin (1894 – 1959). Not long ago, I found a short manuscript of his in which he discussed his views on the place of the theory of probability in the teaching of school mathematics and he noted in general outline the content and nature of presentation.

I am happy that the present little book is accessible to the American reader. During the fifteen years that passed from the time the first Soviet edition was published, many interesting works appeared which extended the field of application of probability theory and about which one could tell in a captivating manner even in a popular booklet. However, I did not wish to disturb the plan or style of what was thought out by my teacher and me in the last months of the war, which swept over the countryside and cities of my country like a hurricane. Changes only touched upon certain examples whose subject matter was determined by the time when the booklet was written. These

changes were made by me in the fifth Soviet edition which is to be published almost simultaneously with the American edition.

24 April 1961. B. V. Gnedenko

Translated from Russian by Leo F. Boron, the translator of the booklet for its American edition

Foreword to the Seventh Edition

For the second time[1] I myself without my teacher and co-author introduce changes by adding a new chapter. When we conceived the ideas of compiling an elementary book on probability theory, we had before our eyes young people graduated from secondary school and thrown away from science by the whirlwinds of the Great Patriotic War. Later, it turned out that the circle of our readers was incomparably wider; it was our book that had acquainted engineers and economists, biologists and linguists, physicians and military men with the ideas and methods of probability theory.

I am pleased that neither in our country nor abroad readers had lost interest in our book. It goes without saying that the change of our readership should somewhat influence the contents of the book. And, since the theory of stochastic processes is now playing a special role in numerous applications of probability theory and in its development, I considered it necessary to add a short introduction to that important field of ideas and studies. Taking into account the general aim of the book, I have accordingly paid most attention to generally acquainting the readers with the practical issues which lead to the theory of those stochastic processes rather than to describing for them the appropriate theory or analytical methods.

I will be really grateful to my readers for submitting any desires concerning the contents or style of the book and the essence of the examples considered there.

Moscow, 10 Dec. 1969. B. V. Gnedenko

From the Foreword to the Eighth Edition

Thirty five years have passed since the appearance of the first edition of this book written on the suggestion of the late Khinchin. After his death I have inserted various changes and additions. The book did not lose readers and I am pleased that some of them have accordingly been led to deep thoughts about applying probabilistic methods in engineering, management and economics.

It is also pleasant that the book had been warmly welcomed abroad; it ran through several editions in [seven countries] and was published in [more than five others]. This edition only differs from the previous by small editorial changes, but life is going on and I would like to hear the readers' wishes about desirable additions and alterations.

Moscow, September 1975. B. V. Gnedenko

Foreword to the Present Translation

I. The book. It has been greatly successful, witness the Forewords to some of its previous editions. To my surprise, it is hardly satisfactory, and I only hope that my appended Notes (unsigned, unlike the authors' Notes now accompanied by letters G&K) and tiny

insertions and question marks in the text itself will explain the situation. Here are my conclusions.

1. The book is written very carelessly as mentioned in 12 of my Notes. Just one example: artillery firing is mentioned more than once, and each time the scatter of the shells is only considered along the line of firing. Only once (Note 35) the authors obliquely remark that the shells *fall around*. Carelessness was apparently the reason for mentioning quite unnecessary details as well. Thus (omitted in the translation), four main causes of stoppages of looms are listed (§ 2.2).

2. Several opportunities to insert important remarks are missed although *the student is a torch to be fired rather than a container to be filled*. The shortcomings of the Bayesian approach are not indicated (Note 13), the Bernoulli theorem is discussed unsatisfactorily (Note 20), chaotic motion is not mentioned (Note 51), direct and inverse theorems are not discussed in a general way (Note 31) and neither is sampling (Note 40). Nothing is said about the required number of significant digits in approximate calculations and the authors themselves mistakenly indicated doubtful and unnecessary digits (Note 11).

3. The notions of probability and expectation are justified by common sense without indicating the accepted formal method; moreover, statistical probability is described as theoretical (Note 27). Possibly confusing additional words (*always*, *purely* random etc.) are inserted into statements and definitions (Note 30).

4. Historical comments are unsatisfactory. Chebyshev is properly mentioned in connection with the law of large numbers, but Poisson is left out (Note 39).

5. Some examples concerning the measurement of distances and artillery firing (Notes 37 and 46) belong to fairyland and the discussion of the errors of measurements (Notes 41 and 42, also Note 37) is unsatisfactory.

6. Population statistics is represented by two examples concerning the sex ratio at birth (carelessly stating the probability of a male birth), see Note 16. The authors should not, however, be blamed for neglecting this field of statistics: millions perished in the GULAG, and the war claimed still more lives. For many years population statistics remained a touchy subject. The results of the census of 1937 were allegedly sabotaged and the Central Statistical Directorate decimated (Sheynin 1998). Kolmogorov (Anonymous 1955, pp. 156 – 158) avoided mentioning population statistics in his report of 1954.

The complete absence of examples based on games of chance seems doubtful.

II. Its American translation of 1961. It is dated since Gnedenko had inserted new additions and even a whole new part (Part 3). Then, the translator, Leo E. Baron, followed the Russian original without any comments and too often left the (naturally, Russian) structure of phrases unaltered. He, or the *Editorial collaborator* Sidney F. Mack, appended a Bibliography but it has no connection with the text itself.

III. The authors. Both are generally known, but I am adding some comments. I published a joint paper with **Gnedenko** (Gnedenko & Sheynin 1978) and certainly know that Gnedenko had successfully

studied the work of Chebyshev, Markov and Liapunov, but that he left aside the history of probability as developed by foreign scholars. This fact is clearly visible here as well as in his essay (Gnedenko 2001 and perhaps before that) which should have appeared 30 years earlier.

Among other methodical and pedagogical contributions **Khinchin** left a concise treatise on mathematical analysis (1948), a possibly rather too shortened textbook for university students (1953) and a posthumously published essay on the Mises theory (1961). Gnedenko edited it and explained that the celebrated journal, *Uspekhi Matematicheskikh Nauk*, had rejected its manuscript. Unfortunately, the cause of rejection remains unknown.

Little known is Kolmogorov's acknowledgement (1933/1956, p. 0003) inserted in his great book:

I wish to express my warm thanks to Mr. Khinchine who has read carefully the whole manuscript and proposed several (*mehrere*!) *improvements*.

On the other hand, Khinchin's invasion of statistical physics (1943) was unsuccessful. Here is Novikov (2002, p. 334) whose paper deserves to be translated in full:

Khinchin attempted to begin studying the justification of statistical physics, but physicists met his contribution on [that subject] *with deep contempt. Leontovich* [an eminent and widely known physicist] *said … that Khinchin does not understand anything*.

But the most disturbing fact is the appearance of Khinchin's (1937) glorification of the Soviet regime published at the peak of the Great Terror. In October 1937 a "Colloque des probabilités" took place at the Genève University. Among the participants were Cramér, Feller, Hostinsky and other most distinguished scholars who signed *Compliments* to Born on the occasion of his birthday (Staatsbibl. Berl. Preussische Kulturbesitz. Manuskriptabt. Nachl. Born 129). No wonder that there were no Soviet participants! Information about the Great Terror should have been prevented. So much for Khinchin's kowtowing …

In 1986, a second edition of the Russian translation of part 4 of Jakob Bernoulli's *Ars Conjectandi* had appeared complete with three commentaries, one of which was mine. A subeditor told me to suppress my reference to Khinchin. He had not elaborated and I, regrettably, did not ask for any explanations. The Editor was the late Yu. V. Prokhorov, a well-known student of Kolmogorov.

Oscar Sheynin

Part 1

Probabilities

Chapter 1. Probabilities

1.1. The Notion of Probability. If under some conditions (the same target, distance and rifle) a certain shot achieves 92% of hits, it follows that on the average he hits the target about 92 times out of a hundred (and therefore fails approximately 8 times). Of course, he will sometimes be successful 91, or 90, or 93 or 94 times, he can even hit the target much less or much more than 92 times, but *in the mean*, after numerous attempts made under the same circumstances, this frequency of hits will remain invariable until there occurs some essential change (for example, this shot can raise his skill and achieve, again in the mean, 95 or more hits out of a hundred).

Experience proves that in most cases shots indeed succeed about 92 times out of a hundred. Less than 88 hits or more than 96 do occur, but only rarely. That figure, 92, the indicator of the skill of our shot, is usually very *stable* which means that under the same conditions the frequency of his hits will in most cases be almost invariable. Only as an exception it will somewhat considerably deviate from its mean value.

One more example. In a certain workshop about 1.6% of the articles manufactured under given conditions are substandard which means that in a batch of, say, a thousand articles, about 16 will be useless. This figure will certainly be sometimes larger and at times smaller, but in the mean it will be near to 16. In most batches of a thousand articles, it will also be near to 16. We certainly suppose that all the conditions of work are invariable.

Such examples can obviously be indefinitely multiplied. And we invariably see that, having *homogeneous mass operations* (when firing many times over, manufacturing articles en masse etc) going on under given conditions, the frequency of one or another occurring important event (of hitting the target, obtaining a substandard article etc) is almost always approximately the same; only rarely does it somewhat considerably deviate from some mean figure.

We may therefore say that, under strictly established conditions, this mean figure is a typical indicator of the given mass operation. The frequency of hits characterizes the skill of the shot; the frequency of the occurred substandard articles estimates the quality of the production. It is thus self-evident that the knowledge of such indications is very important for most various areas: military science, technology, economics, physics, chemistry etc. They allow us to estimate previous mass phenomena as well as to foresee the outcome of some future mass operation.

If on the average and under given conditions a shot hits the target 92 times out of a hundred, we say that for him, and under those conditions, *the probability of hitting the target* amounts to 92% or 92/100 or 0.92. If in a certain workshop, again in the mean and under given conditions, 16 substandard articles occur out of each thousand, we say that *the probability of manufacturing a substandard article* amounts there to 0.016 or 1.6%.

So what do we call the probability of an event in a given mass operation? Now, it is not difficult to answer this question. A mass

operation is always a numerous repetition of similar solitary operations (of shooting, of manufacturing an article etc). We are interested in a certain result of a solitary operation (of a successful single shot, of the quality of a given article etc) but, first of all, in the number of such results in some mass operation (in the number of hits, of substandard articles etc).

The relative frequency of a *successful* result[2] in a given mass operation we will indeed call tits *probability*. However, we should always bear in mind that the probability of one or another event (result) only makes sense if our mass operation goes on under strictly defined conditions. As a rule, any essential change of those conditions leads to a change of the appropriate probability.

Suppose that in the mean an event *A* (for example, a successful hit of the target) is achieved *a* times out of *b* solitary operations (shots) of a mass operation. Then the probability of a *successful* outcome of a solitary operation is *a/b, the ratio of the mean achieved number of such successful outcomes to the number of all the solitary operations* comprising the given mass operation.

If the probability of some event is *a/b*, it can obviously appear either more or less often than *a* times in each series of *b* solitary operations. Indeed, it only occurs about *a* times in the mean and in most series of *b* operations the number of the occurrences of that event will be near to *a, especially if b is a large number*.

Example 1. In a certain town there were born during the first quarter of some year:

In January, 145 boys (b) and 135 girls (g); in the next two months, respectively, 142 b and 136 g; and 152 b and 140 g. How high is the probability of a male birth? The relative frequencies were

$$145/280 \approx 0.518 = 51.8\%;\ 142/278 \approx 0.511 = 51.1\%;$$
$$152/292 \approx 0.521 = 52.1\%$$

The arithmetic mean of those frequencies is near to 0.516 = 51.6%. Under given conditions it is the probability sought and the figure 0.516 is well known in demography, the science that studies changes in populations. Under usual conditions the relative frequencies of male births during different intervals of time do not essentially deviate from that figure.

Example 2. A remarkable phenomenon was discovered in the beginning of the 19th century: *the Brownian motion* named after the British botanist Brown. Minutest particles of a substance suspended in a liquid[3] are moving chaotically and without any visible reason. For a long time the cause of this apparently spontaneous motion remained unknown, but the kinetic theory of gases provided a simple and exhaustive explanation. That motion is due to the shocks inflicted by the molecules of the liquid upon those particles. The kinetic theory allows us to calculate the probabilities that in a given volume of the liquid there will be none, one, two, … particles of the suspended substance.

Series of experiments had been carried out for checking the theoretical results. We provide the data of 518 observations of

minutest particles of gold suspended in water made by the Swedish physicist Svanberg. He found out that in the observed space no particles occurred in 112 cases; 1 particle appeared 168 times; and 2, 3, 4, 5, 6 and 7 particles appeared 130, 69, 32, 5, 1 and 1 time (times). The relative frequencies of those cases were

112/518 = 0.216; 168/518 = 0.324; 130/518 = 0.251;
69/518 = 0.133; 32/518 = 0.062; 5/518 = 0.010; and
(twice) 1/518 = 0.002

His results agreed very well with the theoretically predicted probabilities.

Example 3. In a number of practically important problems it is very important to know the possible relative frequencies of the occurrence of different letters of the Russian [Cyrillic] alphabet. Thus, when compiling a set of types in a printing house it is impractical to collect the same number of each letter since some occur considerably oftener than the others. Studies of literary texts led to the estimation of those frequencies, see Table 1 borrowed from A. M. and I. M. Yaglom (1957/2007).

[Table 1 lists the relative frequencies of the space between words and of the 31 letter of the alphabet (letters *e* and *ë* are combined). Some frequencies: the space, 0.175, letters *o* and (*e* and *ë*): 0.090 and 0.072, then down to *э* and *ф*: both 0.002.] Thus, out of a thousand randomly chosen spaces and letters *ф* will occur twice; letters *к* and *o* and the space will be found 28, 90 and 175 times. The Table provides a sufficiently valuable indication for compiling sets of types.

Such investigations have recently been also widely applied for revealing peculiar features of the Russian language and of the literary style of various authors. Information about wire messages can be applied for constructing optimal wire codes and thus ensuring a faster transmission of messages by a lesser number of signs. It turned out that the wire codes which had been applied during World War II were not sufficiently economical[4].

1.2. Impossible and Certain Events. The probability of an event is always either a positive number or zero. It cannot exceed 1 since the numerator of the fraction that determines it cannot be larger than its denominator (the number of *successful* operations cannot exceed the number of all of them). We will denote the probability of event *A* by P(*A*) so that for any event

$$0 \le P(A) \le 1.$$

The higher is P(*A*) the more often will event *A* occur. Thus, the higher is the probability of hitting the target, the more often will the shot achieve his goal, the more skilful he is. If, however, the probability of an event is very low, it occurs rarely, and if P(*A*) = 0, event *A* either never occurs or happens extremely rarely and can therefore be considered *practically impossible*. On the contrary, if P(*A*) is near to unity, the numerator of the fraction which expresses that

probability is near to its denominator, an overwhelming number of operations is successful and such events occur in most cases.

If P(A) = 1, event A appears always or almost always and we may consider it *practically certain*, and reckon on its occurrence for sure. Then, if P(A) = 1/2, event A happens in about a half of the cases so that approximately the *successes* and *failures* are equally numerous. If P(A) > 1/2, event A occurs more often than fails; otherwise, if P(A) < 1/2, it happens less often than fails.

How low should the probability of an event be for considering it practically impossible? No all-embracing answer is possible, all depends on how important is the problem at hand. Thus, 0.01 is a small number. If it is the probability that a shell will not explode on impact, about 1% of the fired shells will be useless. This can be accepted, but the same probability that a parachute will not open is certainly intolerable.

These examples prove that for each problem the admissible low probability of an event ought to be established beforehand for harmlessly deciding that that event is practically impossible.

1.3. A Problem. A shot hits the target with probability 0.80; another shot, under the same conditions, hits it with probability 0.70. Both fire at the same time and required is the probability of hitting the target at least once.

Solution, first method. Suppose that they fire 100 times. The first shot will hit the target approximately 80 times and fail about 20 times. The second shot succeeds 70 times in the mean or 14 times when the first shot fails. The target is hit 80 + 14 = 94 times and the probability of success, when both are firing at the same time, is 94% or 0.94.

Solution, second method. Again suppose that they fire a hundred times. The first shot fails about 20 times, the second shot fails approximately 30 times, or about 3 times in 10 [6 times in 20]. It can therefore be expected that 6 times the target will not be hit by either, but 94 times at least once. The result is the same as above.

This problem is very easy, but nevertheless it leads to a very important conclusion: There are cases in which the probabilities of more complicated events can be expediently derived from the probabilities of simple or less complicated events. Actually, there are very many such cases, and not only in military science but in any science, in any practical activity in which mass phenomena are involved.

It would certainly be inconvenient to devise a special method for solving each new problem; science invariably attempts to create general rules for solving similar problems mechanically or almost so. In the area of mass phenomena the science that shoulders the compilation of such rules is called *the theory of probability*, and here we offer its elements[5]. It is a chapter of mathematics just like arithmetic or geometry. Consequently, it works by strict reasoning and its tools are formulas, tables, charts etc.

Chapter 2. The Rule for the Addition of Probabilities

2.1. The Derivation of the Addition Rule. This is the simplest and the most important rule applied for calculating probabilities. For each shot firing at a target, there exists some probability of hitting it from a given distance. Let 1 be a small circle drawn on the target, and denote concentric circles of increasing radii forming rings by 2, 3, 4 and 5 and the region partly situated beyond the target by 6. Suppose now that a certain shot has probability 0.24 of hitting circle 1, and of hitting ring 2, 0. 17. We already know that in the mean 24 of his bullets out of a hundred will hit circle 1, and 17 bullets will hit ring 2. Call such results *excellent* and *good* and determine the probability that the result of one attempt will be either excellent or good.

This is an easy problem. Approximately 24 and 17 bullets out of a hundred will ensure excellent and good results respectively, so that 41 bullet will hit either the circle or the ring. The probability sought will be 0.24 + 0.17 = 0.41. It is thus the sum of the probabilities of an excellent and a good result.

Another example. A man is awaiting either tram 26 or tram 16. He is standing at a tram stop for trams 16, 22, 26 and 31, and we suppose that all of them come approximately equally often. What is the probability that the first tram to appear will be of the required route? The probability that the first tram will be number 16 is obviously 1/4, and the same probability exists for tram 26. The probability sought is therefore 1/2, the sum of 1/4 and 1/4, of the two probabilities.

And now the general reasoning concerning some mass operation. Suppose that it is established that on average in each series of b solitary operations

a_1 times result A_1 was observed, a_2 times result A_2, etc.

In other words, the probability of event A_1 is a_1/b, of event A_2, a_2/b, etc. Required is the probability that any one of those results will occur in a solitary operation. The event which interests us can be denoted by (A_1 or A_2 or …). In a series of b operations it occurs ($a_1 + a_2 + \ldots$) times and the probability sought is

$$\frac{a_1 + a_2 + \ldots}{b} = \frac{a_1}{b} + \frac{a_2}{b} + \ldots$$

This can be written as

$$P(A_1 \text{ or } A_2 \text{ or } \ldots) = P(A_1) + P(A_2) + \ldots$$

In both examples and here, in the general reasoning, we assumed that any two of the considered events (for example, A_1 and A_2) are *mutually incompatible*, i.e., that they cannot occur in the same solitary operation. Thus, a tram cannot be both needed and not needed, it either satisfies the passenger's need or not.

This assumption about mutual exclusiveness of the separate results is very important. If it does not take place, the addition rule stated

below becomes wrong and its application leads to gross mistakes. Consider for example the problem solved at the end of § 1.3. It was required to derive the probability that, when both shots fire at the same time, the target will be hit either by the first of them or by the second. Their probabilities of success were 0.8 and 0.7 and a direct application of the addition rule leads to probability of success equal to 0.8 + 0.7 = 1.5. This is nonsense since probabilities cannot exceed unity. We arrived at this wrong and senseless result since we used the addition rule in a case in which it is inapplicable: those two individual results are *compatible*. It is quite possible that both shots hit the target at the same time.

A considerable part of mistakes made by beginners when calculating probabilities is caused by such a wrong application of the addition rule. It is therefore necessary to be invariably on guard against this mistake. When applying the addition rule, check without fail whether each two of the studied events are really incompatible.

And now we may formulate the general *addition rule*:

The probability of the occurrence in some operation of any of the results A_1, A_2, …, A_n (no matter which) is equal to the sum of their probabilities if only each two of them are mutually incompatible.

2.2. Complete Systems of Events. One third of a certain State loan bonds gradually wins during a twenty-year period, the other bonds are then repaid. Each bond thus wins with probability 1/3 and is repaid with probability 2/3. The two events are *contrary* which means that one and only one of them certainly occurs. The sum of their probabilities, 1/3 + 2/3, is unity which is not accidental.

In general, if A_1 and A_2 are contrary events, and in a series of b operations A_1 occurs a_1 times, and A_2 occurs a_2 times, then, obviously, $a_1 + a_2 = b$. And indeed,

$$P(A_1) = a_1/b,\ P(A_2) = a_2/b,$$
$$P(A_1) + P(A_2) = a_1/b + a_2/b = (a_1 + a_2)/b = 1.$$

The addition rule leads to the same result: since contrary events are mutually incompatible,

$$P(A_1) + P(A_2) = P(A_1 \text{ or } A_2).$$

But the event (A_1 or A_2) is certain since it ought to occur by definition of contrary events and its probability is unity. We again have

$$P(A_1) + P(A_2) = 1.$$

The sum of the probabilities of two contrary events is unity.

This rule can be very importantly generalized which is proved in a similar way. Suppose that there are such arbitrarily many (n) events A_1, A_2, …, A_n that one and only one of them occurs without fail in each solitary operation. Such group of events is called a *complete system*. In particular, any pair of contrary events makes up a complete system.

The sum of the probabilities of the events comprising a complete system is unity.

Indeed, by the definition of a complete system any two of its events are mutually incompatible and

$$P(A_1) + P(A_2) + \ldots + P(A_n) = P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n).$$

The right side of this equality is the probability of a certain event and therefore equals unity:

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1, \text{ QED.}$$

Example 1. A shot firing at a target described in § 2.1 hits, in the mean, 44 times the circle 1; 30, 15, 6, 4 and 1 hit-point (points) is (are) contained within rings 2, 3, 4, 5 and region 6, and

$$44 + 30 + 15 + 6 + 1 = 100.$$

These results obviously constitute a complete system of events whose probabilities are 0.44, 0.30, 0.15, 0.06, 0.04 and 0.01. Their sum is unity. Some hit-points in region 6 cannot be counted, but the appropriate probability can be calculated by subtracting the sum of all other probabilities from unity.

Example 2. At a certain factory out of each hundred stoppages of a loom which required the weaver's attention, 22, 31, 27, and 3 in the mean occurred because of four definite main causes respectively (and the other stoppages had other causes). The probabilities of those four causes are 0.22, 0.31, 0.27 and 0.03. Their sum is 0.83 and the probability of the other causes is $1 - 0.83 = 0.17$ since all the causes comprise a complete system of events.

2.3. Examples. The theorem about the complete system of events often successfully serves as a foundation for the so called *prior* calculation of probabilities. Suppose that we study the distribution of cosmic particles over a rectangular surface subdivided into 6 identical squares. We have no sufficient reason to assume that those particles will more often come to rest on one of these squares rather than on another[6].

So let us assume that the appropriate probabilities $p_1, p_2, \ldots, p_5, p_6$ are identical. If we are only interested in observing the rectangular surface, each of the six probabilities will be equal to 1/6 since their sum is unity by the theorem above. This conclusion depends on assumptions and ought to be verified by experiments, but in such cases we are so accustomed to a positive answer that are practically fully justified to rely on those theoretical assumptions even before experimentation.

In such cases we usually say that the given operation can have n different *equally probable* results; here, the cosmic particles can come to rest on any of the six squares with probability 1/6. Such prior calculations are important since they allow us to foresee events when mass operations are either impossible or extremely difficult to carry out.

Example 1. The numbers of Soviet State bonds had usually been expressed by five digits (for example, 59607). Required is the probability that the last digit of a randomly chosen winning bond is 7.

According to the definition of probability, we have to consider a long table of drawings and find out in how many cases the last digit of the number of winning bonds was seven. The ratio of the number of those cases to the complete number of winning bonds will be the probability sought. However, each of the 10 digits 0, 1, 2, …, 8, 9 has the same chance to be the last one of the number of the winning bond. Without hesitating, we assume that the probability sought is 0.1. Verification of this *prevision* is easy: select the table of any drawing and convince yourself in that each of the 10 digits is the last one in the numbers of the winning bonds in approximately 1/10 of all cases[7].

Example 2. A telephone line between points A and B two kilometres apart was torn somewhere. Required is the probability that the rupture occurred not farther than 450 *m* from A. So let us mentally separate the entire distance AB into metres. Since all the appeared intervals are practically homogeneous, we may assume that the probabilities of the rupture occurring in each of them are identical. Similar to the above, we easily decide that the probability sought is 450/2000 = 0.225.

Chapter 3. Conditional Probabilities and the Multiplication Rule

3.1. The Notion of Conditional Probability. Suppose that two factories are manufacturing light bulbs and that they produce 70 and 30% of the entire output respectively with standard bulbs[8] comprising 83 and only 63% of their respective totals. It is not difficult to calculate that in the mean the customer gets 77 standard bulbs out of a hundred or that the probability of buying a standard bulb is 0.77. Indeed, $0.7 \cdot 83 + 0.3 \cdot 63 = 77$. But suppose now that you buy bulbs only manufactured by the first factory, then that probability will be 0.83.

This example shows that when the general conditions under which an operation is proceeding are coupled with an essentially new condition (only the bulbs produced by the first factory are taken into account) the probability of one or another outcome of a solitary operation can change. This is evident since the notion itself of probability requires that the set of conditions, under which the given mass operation is going on, is strictly established. Generally, an addition of some new condition to that set essentially changes it. The mass operation will continue under new conditions, it will be another operation and consequently the probabilities of one or another ensuing result will change.

And so, we have two differing probabilities of the same event, of a purchase of a standard bulb. Until we impose an additional condition (until taking into account the manufacturer) we have an *absolute probability*, 0.77, of buying a standard bulb; after adding the new condition we have a *conditional probability* of the same event, 0.83, somewhat differing from 0.77.

Denote the event of purchasing a standard bulb by A, and let B be its being manufactured by the first factory. Then $P(A)$ usually denotes the absolute probability of A, and $P_B(A)$, the probability of the same event provided that the bulb was manufactured by the first factory. Then

$$P(A) = 0.77,\ P_B(A) = 0.83.$$

Since probabilities, strictly speaking, are only defined under some exactly determined conditions, then, in the literal sense, each probability is conditional whereas absolute probabilities do not exist. However, in most problems all the studied operations are proceeding under a definite set of conditions K which are supposed to be invariably satisfied. When no other conditions except K are imposed, we call the ensuing probability *absolute*, and *conditional* if in addition some other strictly stipulated condition(s) is (are) imposed.

Thus, in our example we certainly assume that bulbs are manufactured under some definite conditions, invariable for all of them. This assumption is so unavoidable and self-evident that we did not even deem it necessary to mention it. If, when purchasing a bulb, we do not impose any additional condition, the probability of one or another result of its test is called absolute. When some additional conditions are imposed, the appropriate probabilities will be conditional.

Example 1. Above, the probability of a bulb being manufactured by the second factory is obviously 0.3. It is established that the bulb is standard and required is the probability that it was manufactured by that second factory.

Out of every thousand bulbs put on sale 770 are on the average standard, 581 of them manufactured by the first factory, and 189, by the second[9]. The probability sought is $189/770 \approx 0.245$. This is the conditional probability calculated under the assumption that the bulb is standard, and we may conclude that

$$P(\bar{B}) = 0.3,\ P_A(\bar{B}) \approx 0.245$$

where $\bar{B}$ denotes the failure of event *B*.

Example 2. Observations made during many years in a certain district established that, out of 100,000 children aged 10, 82,277 in the mean live until 40 years and 37,977, until 70 years. Required is the probability that a man **[the sexes are not distinguished]** 40 years old will live until 70 as well. The probability sought is $37,977/82,277 \approx 0.46$.

Denote by *A* and *B* the events of a child 10 years old living until 70 and 40 years. Then, obviously,

$$P(A) = 0.37977 \approx 0.38,\ P_B(A) \approx 0.46.$$

3.2. Derivation of the Rule for Multiplying Probabilities. Return to the first example of § 3.1. Out of a thousand bulbs put on sale the second factory manufactures 300, 189 of them in the mean being standard. The probability that the bulb was manufactured by the second factory (event $\bar{B}$) is

$$P(\bar{B}) = 300/1000 = 0.3$$

and the probability of its being standard given that it was manufactured by the second factory is

$$P_{\bar{B}}(A) = 189/300 = 0.63.$$

And so, out of the 1000 bulbs 189 are manufactured by the second factory and are standard and the probability of the joint occurrence of events *A* and $\bar{B}$ is

$$P(A \text{ and } \bar{B}) = 189/1000 = (300/1000)(189/300) = P(\bar{B})\,P_{\bar{B}}(A).$$

This *multiplication rule* can easily be extended on the general case. Suppose that result *B* occurs *m* times in the mean in each series of *n* operations, and in each new series of *m* such operations *l* times appears result *A*. Then this joint occurrence of events *B* and *A* in each series of *n* operations will occur in the mean *l* times. Thus,

$$P(B) = m/n,\ P_B(A) = l/m,\ P(A \text{ and } B) = l/n = (m/n)(l/m) = P(B)P_B(A). \quad (3.1)$$

The multiplication rule. *The probability of a joint occurrence of two events is equal to the product of the probability of the first event by the conditional probability of the second calculated under the assumption that the first event had occurred*.

We may certainly say that any of the two events is the first one so that in addition to formula (3.1) we may just as well say that

$$P(A \text{ and } B) = P(A)P_A(B). \qquad (3.1^*)$$

An important relation follows:

$$P(A)P_A(B) = P(B)P_B(A). \qquad (3.2)$$

In our example we had

$$P(A \text{ and } B) = 189/1000,\ P(A) = 77/100,\ P_A(\bar{B}) = 189/770$$

and formula (3.1*) is confirmed.

Example. 96% of the articles manufactured at a certain factory are suitable (event *A*) and 75% of them are of top quality (event *B*). Required is the probability that an article manufactured there is of top quality. We ought to find P(*A* and *B*) since a top quality article (event *B*) should first of all be suitable (event *A*).

According to the conditions of the problem

$$P(A) = 0.96,\ P_A(B) = 0.75$$

and by formula (3.1*)

$$P(A \text{ and } B) = 0.96 \cdot 0.75 = 0.72.$$

3.3. Independent Events. After a test of tensile strength of two skeins of thread produced by different looms it occurred that a specimen of some length taken from the first skein endures a certain standard load with probability 0.84, and with probability 0.77 if taken from the second skein[10]. Required is the probability that both these specimens endure that load (event *A* and *B*).

We require P(*A* and *B*) and apply the multiplication rule

$$P(A \text{ and } B) = P(A)P_A(B).$$

Here $P(A) = 0.84$, but what is the meaning of $P_A(B)$? According to the general definition of conditional probabilities, it is the probability that the specimen taken from the second skein endures the load if the specimen from the first skein endures it. However, the probability of *B* does not depend on event *A* occurring or not. Practically speaking, it means that the per cent of tests in which *B* takes place does not depend on the strength of the specimen taken from the first skein[11]:

$$P_A(B) = P(B) = 0.78,\ P(A \text{ and } B) = P(A)P(B) = 0.84 \cdot 0.78 = 0.6552.$$

As compared with all the previous examples, this one is peculiar in that, as we see, the probability of *B* does not change when we add to the general conditions the requirement that *A* ought to occur. In other words, the conditional probability $P_A(B)$ is equal to the absolute probability P(*B*). In this case we simply say that *event B does not depend on event A*.

It is easy to check that in this case *A* does not depend on *B* either. Indeed, since $P_A(B) = P(B)$, then, by formula (3.2), $P_B(A) = P(A)$ as well which indeed means that event *A* does not depend on event *B*. Independence of two events is thus a mutual property. We see that for independent events the multiplication rule becomes especially simple:

$$P(A \text{ and } B) = P(A)P(B). \qquad (3.3)$$

Whenever we apply the addition rule we ought to establish beforehand that the given events are incompatible. Just the same, whenever we apply the rule (3.3) we ought to establish whether the events *A* and *B* are independent. Neglect of this indication leads to a large number of mistakes. If the events *A* and *B* are dependent, formula (3.3) becomes wrong and should be substituted by a more general formula (3.1) or (3.1*).

Rule (3.3) is easily generalized on the probability of three or more independent events. Suppose we have three *mutually independent* events *A, B* and *C*. This means that the probability of neither of them depends on whether the other events occurred or not. And since the three events are independent, according to formula (3.3)

$$P(A \text{ and } B \text{ and } C) = P(A \text{ and } B)P(C).$$

Again apply formula (3.3) for determining P(*A* and *B*), then

$$P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C). \qquad (3.4)$$

The same rule obviously takes place when the studied group consists of any number of events if only they are *mutually independent*, if the probability of each does not depend on whether the other events occurred or not.

The probability of the joint occurrence of any number of mutually independent events is equal to the product of their probabilities.

Example 1. A worker services three lathes. The probabilities that no service is needed for an hour are 0.9, 0.8 and 0.85 respectively. Required is the probability that during an hour service will not be needed at all.

Assume that the lathes are operating independently from each other. Then, by formula (3.4), the probability sought is

$$0.9 \cdot 0.8 \cdot 0.85 = 0.612.$$

Example 2. Retain the conditions of the previous example and determine the probability that during an hour at least one lathe will not require attention. The probability sought is of the type P(*A* or *B* or *C*)

and we certainly begin thinking about the addition rule. However, we see at once that that rule is here inapplicable: any two events, and even all three of them are compatible with each other. Indeed, two or even three lathes can certainly continue working during an hour. And even independently from this consideration we immediately see that the sum of the three given probabilities considerably exceeds unity and cannot therefore be any probability at all.

For solving this problem we note that the contrary probabilities of the lathes requiring attention are 0.1, 0.2 and 0.15. They are mutually independent and by rule (3.4) the probability that all of them occur is

$$0.1 \cdot 0.2 \cdot 0.15 = 0.003.$$

However, the events *all three lathes require attention* and *at least one does not require attention* are contrary, their sum is unity and the probability sought is therefore $1 - 0.003 = 0.997$. When a probability of an event is so high, we may assume that it is practically certain. This means that during an hour at least one lathe will continue working.

Example 3. 250 devices are tested under specific conditions. The probability that a definite device fails during an hour is 0.004, the same for all of them. Required is the probability that during an hour at least one device fails.

For one device the probability of working during an hour is $1 - 0.004 = 0.996$ and the probability that none fails is, by the multiplication rule for mutually independent events, 0.996^{250}. The probability sought is $1 - 0.996^{250} \approx 5/8$.

Although the probability of a failure for each device is not high **[is tiny]**, for a large number of them the probability of at least one failure is rather considerable. The reasoning in the two last examples can be easily generalized and lead to an important general rule. In both cases we discussed probabilities $P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n)$ of the occurrence of at least one of some mutually independent events $A_1, A_2, \dots, A_n$. Denote by $\overline{A}_k$ the failure of event A_k, then A_k and $\overline{A}_k$ are contrary and

$$P(A_k) + P(\overline{A}_k) = 1.$$

On the other hand, events $\overline{A}_1, \overline{A}_2, \dots, \overline{A}_n$ are obviously independent so that

$$P(\overline{A}_1 \text{ and } \overline{A}_2 \text{ and } \dots \text{ and } \overline{A}_n) =$$
$$P(\overline{A}_1)P(\overline{A}_2) \dots P(\overline{A}_n) = [1 - P(A_1)]\,[1 - P(A_2)] \dots [1 - P(A_n)].$$

Finally, events $(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n)$ and $(\overline{A}_1 \text{ and } \overline{A}_2 \text{ and } \dots \text{ and } \overline{A}_n)$ are obviously contrary (either at least one event A_k occurs or all the events $\overline{A}_k$ take place). Therefore,

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n) = 1 - P(\overline{A}_1 \text{ and } \overline{A}_2 \text{ and } \dots \text{ and } \overline{A}_n) =$$
$$[1 - P(A_1)]\,[1 - P(A_2)] \dots [1 - P(A_n)]. \qquad (3.5)$$

This important formula allows us to calculate the probability of the occurrence of *at least* one of the events $A_1, A_2, \ldots, A_n$ given their probabilities. It is valid then and only then when all those events are mutually independent. In particular, when all the events A_k have the same probability p (as in Example 3),

$$\mathrm{P}(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = 1 - (1 - p)^n. \qquad (3.6)$$

Example 4. A machine part is a right parallelepiped. It is suitable if the length of each of its edges deviates from the standard not more than by 0.01 *mm*. The probabilities of such unacceptable deviations of lengths, widths and heights are

$p_1 = 0.08$, $p_2 = 0.12$, $p_3 = 0.1$.

Required is the probability that a machine part is unsuitable. This happens when at least one deviation exceeds 0.01 *mm*. Deviations of the three dimensions are usually considered mutually independent (since they are occasioned by different causes) and we may therefore apply formula (3.5):

$$1 - (1 - p_1)(1 - p_2)(1 - p_3) \approx 0.27.$$

Out of each hundred machine parts 73 in the mean will be suitable.

Chapter 4. Corollaries of the Addition and Multiplication Rules

4.1. Derivation of Some Inequalities. Return to our example concerning light bulbs (§ 3.1) once more and denote

A, a standard bulb; $\overline{A}$, a substandard bulb;

B and $\overline{B}$, a bulb manufactured by the first and the second factory

Events *A* and $\overline{A}$ are obviously contrary just as events *B* and $\overline{B}$. If a bulb is standard (event *A*), it is manufactured either by the first (event *A* and *B*) or by the second (event *A* and $\overline{B}$) factory. These two events are obviously incompatible and by the addition rule

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \overline{B}), \quad (4.1)$$

$$P(B) = P(A \text{ and } B) + P(\overline{A} \text{ and } B). \quad (4.2)$$

Consider now event (*A* or *B*). There are three possibilities for its occurrence: *A* and *B*; *A* and $\overline{B}$; and $\overline{A}$ and *B*. Any two of them are incompatible and by the addition rule we have

$$P(A \text{ or } B) = P(A \text{ and } B) + P(A \text{ and } \overline{B}) + P(\overline{A} \text{ and } B). \quad (4.3)$$

Adding up equations (4.1) and (4.2) and taking into account equation (4.3) we easily derive

$$P(A) + P(B) = P(A \text{ and } B) + P(A \text{ or } B),$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \quad (4.4)$$

This is a very important result. True, we considered a particular example, but our reasoning was so general that the conclusion can be thought to hold for any pair of events *A* and *B*. Until now, we only derived expressions for the probability P(*A* or *B*) under very particular assumptions about the connection between those events, *A* and *B* (at first, we considered them incompatible, then mutually independent).

However, formula (4.4) takes place without any additional assumptions for any pair of events *A* and *B*. True, we should not forget an essential difference between formula (4.4) and our previous formulas. Until now, the probability P(*A* or *B*) had always been only expressed through P(*A*) and P(*B*) and we were then invariably able to derive a single value for the event (*A* or *B*).

The essence of formula (4.4) differs: in addition, we have to know P(*A* and *B*), the probability of the joint occurrence of the events *A* and *B*. In the general case, in which the connection between these events is arbitrary, it is usually not easier to calculate that probability than the probability P(*A* or *B*). Consequently, formula (4.4) is rarely applied although its theoretical importance is very considerable.

And now, by issuing from it, we will easily derive our previous formulas. If events *A* and *B* are incompatible, the event (*A* and *B*) is

impossible, P(A and B) = 0, and formula (4.4) leads to the addition formula

$$P(A \text{ and } B) = P(A) + P(B).$$

Then, if A and B are independent, formula (3.3) provides

$$P(A \text{ and } B) = P(A)P(B)$$

and formula (4.4) leads to

$$P(A \text{ or } B) = P(A) + P(B) - P(A)P(B) = 1 - [1 - P(A)][1 - P(B)]$$

which is formula (3.5) for $n = 2$.

We will now derive an important corollary of the same formula (4.4). Since identically P(A and B) ≥ 0, it follows that always

$$P(A \text{ or } B) \leq P(A) + P(B). \qquad (4.5)$$

This inequality can be generalized on any number of events. Thus, for three events,

$$P(A \text{ or } B \text{ or } C) \leq P(A \text{ or } B) + P(C) \leq P(A) + P(B) + P(C)$$

and we can now pass on to four events etc. Here is the general result:

The probability of the occurrence of at least one event out of some number of them is never higher than the sum of their probabilities.

Equality only takes place when any two of those events are incompatible.

4.2. The Formula for Complete Probability. Return once more to our example concerning light bulbs (§ 3.1) and to our notation in § 4.1. We have more than once seen that the probabilities that a standard bulb was manufactured by the second and the first factories were

$$P_{\bar{B}}(A) = 189/300 = 0.63,\ P_B(A) = 581/700 = 0.83.$$

Suppose that both these probabilities are known as well as the probabilities that the bulb was manufactured by those factories **[for some reason mentioned in the inverted order: by the first and second factories]**

$$P(B) = 0.7,\ P(\bar{B}) = 0.3.$$

Required is the absolute probability P(A) that a bulb is standard without any assumptions about its manufacturer. Let us reason in the following way. Denote by E and F the compound events that the bulb was manufactured by the first and the second factory. Each bulb was manufactured either by the first or the second factory and therefore event A is tantamount to event (E or F). These events are incompatible and by the addition rule

$$P(A) = P(E) + P(F). \qquad (4.6)$$

On the other hand, as it follows from the above, event E is tantamount to event (A and B). Therefore, according to the multiplication rule,

$$P(E) = P(B)P_B(A)$$

and in exactly the same way

$$P(F) = P(\bar{B})P_{\bar{B}}(A).$$

Substituting these two expressions in equality (4.6) we obtain

$$P(A) = P(B)P_B(A) + P(\bar{B})P_{\bar{B}}(A),$$

the formula that solves our problem. Substituting the given data we get $P(A) = 0.77$.

Example. Seeds of wheat of sort I are stored for sowing. They contain a small admixture of seeds of sorts II, III, and IV. Choose a seed and denote its being of these sorts by A_1, A_2, A_3 and A_4 respectively. It is known that

$$P(A_1) = 0.96,\ P(A_2) = 0.01,\ P(A_3) = 0.02 \text{ and } P(A_4) = 0.01.$$

The sum of these probabilities is unity as it should be for a complete system of events. The probabilities that an ear containing not less than 50 grains will grow out of a seed are respectively 0.50, 0.15, 0.20 and 0.05. Required is the absolute probability of an ear having not less than 50 grains (event K).

By the conditions of the problem

$$P_{A_1}(K) = 0.50,\ P_{A_2}(K) = 0.15,\ P_{A_3}(K) = 0.20,\ P_{A_4}(K) = 0.05.$$

The probability sought is $P(K)$. Denote by E_1 the event that a seed is of sort I and that the ear which grew out of it has not less than 50 grains. That event, E_1, is therefore tantamount to event (A_1 and K). We also denote by E_2, E_3 and E_4 similar events (A_2 and K), (A_3 and K) and (A_4 and K).

For event K to arrive the occurrence of one of the events E_1, E_2, E_3 or E_4 is necessary. Since any two of them are incompatible, the addition rule provides

$$P(K) = P(E_1) + P(E_2) + P(E_3) + P(E_4). \qquad (4.7)$$

On the other hand, by the multiplication rule

$$P(E_1) = P(A_1 \text{ and } K) = P(A_1)P_{A_1}(K)$$

and similar expressions hold for E_2, E_3 or E_4. Substituting these expressions into formula (4.7) we get

$$\mathrm{P}(K) = \mathrm{P}(A_1)\,\mathrm{P}_{A_1}(K) + \mathrm{P}(A_2)\,\mathrm{P}_{A_2}(K) + \mathrm{P}(A_3)\,\mathrm{P}_{A_3}(K) + \mathrm{P}(A_4)\,\mathrm{P}_{A_4}(K), \tag{4.8}$$

a formula that evidently answers our problem. Substituting the given data we obtain $\mathrm{P}(K) = 0.486$.

The examples considered here in detail lead to an important general rule which we are now able to formulate and justify without any difficulties. Suppose that a given operation admits results A_1, A_2, …, A_n constituting a complete system of events. To recall: it means that any two of those events are incompatible with each other and one of them occurs for sure.

Then for any possible result K of this operation to occur we have formula (4.8) with n terms instead of 4. It is usually called *the formula of complete probability*. It is proved just like it was done in the two examples above. First, the appearance of event K requires the occurrence of one of the events (A_i and K), and, by the addition rule,

$$\mathrm{P}(K) = \sum_{i=1}^{n} \mathrm{P}\left(A_i \text{ and } K\right). \tag{4.9}$$

Second, according to the multiplication rule,

$$\mathrm{P}(A_i \text{ and } K) = \mathrm{P}(A_i)\,\mathrm{P}_{A_i}(K).$$

Substituting this expression in equality (4.9) we will indeed arrive at formula (4.8).

4.3. The Bayes Formula. The formulas of § 4.2 allow us to derive an important result having numerous applications. We begin by a formal justification and postpone the ascertaining of the meaning of the final formula.

Suppose that once more events A_1, A_2, …, A_n constitute a complete group of the results of some operation. If K is one of these results, then, by the multiplication rule,

$$\mathrm{P}(A_i \text{ and } K) = \mathrm{P}(A_i)\,\mathrm{P}_{A_i}(K) = \mathrm{P}(K)\mathrm{P}_K(A_i),\ 1 \le i \le n$$

and therefore

$$\mathrm{P}_K(A_i) = \frac{\mathrm{P}(A_i)\mathrm{P}_{A_i}(K)}{\mathrm{P}(K)},\ 1 \le i \le n.$$

Expressing the denominator by the formula of complete probability (4.8) we get *the Bayes formula*

$$P_K(A_i) = \frac{P(A_i)P_{A_i}(K)}{\sum_{r=1}^{n} P(A_r)P_{A_r}(K)},\ 1 \le i \le n. \qquad (4.10)$$

This formula is mostly applied as is shown in the following example. Suppose that a target is situated on a segment MN which we mentally separate into five small intervals c_2, b_2, a, b_1, c_1 in that order. The exact location of the target is unknown and we can only say that the probabilities of its being on those intervals are

$$P(a) = 0.48,\ P(b_1) = P(b_2) = 0.21,\ P(c_1) = P(c_2) = 0.05.$$

The sum of these probabilities is unity. The highest probability corresponds to interval *a*, and we naturally fire at it. However, owing to unavoidable errors the target can also be hit if it is located elsewhere. The probabilities of the hits are $P_a(K) = 0.56$ if the target is located on a. Other probabilities are

$$P_{b_1}(K) = 0.18,\ P_{b_2}(K) = 0.16,\ P_{c_1}(K) = 0.06,\ P_{c_2}(K) = 0.02$$

[the sum of these probabilities is 0.98].

Suppose now that the target is hit (event *K* has occurred). The probabilities of the location of the target, i. e., the numbers P(a), $P(b_1)$, …, are estimated anew[12]. The qualitative essence of this operation is evident even without any calculations. If we aimed at interval a and hit the target, the probability P(a) ought to be heightened.

However, we wish to estimate numerically this reappraisal, to derive an exact expression of the probabilities $P_K(a)$, $P_K(b_1)$, …given that the target was hit. The Bayes formula (4.10) immediately provides the answer:

$$P_K(a) = \frac{P(a)P_a(K)}{P(a)P_a(K) + P(b_1)P_{b_1}(K) + \ldots + P(c_2)P_{c_2}(K)} \approx 0.8.$$

We see that $P_K(a)$ is indeed higher than P(a).

In a similar way we can easily calculate the probabilities $P_K(b_1)$, … of the other possible locations of the target. When actually calculating, it is expedient to note that the expression of the probabilities as provided by the Bayes formula only differ in their numerators; the denominators are the same and equal to $P(K) \approx 0.34$.

We can describe the general pattern in the following way. The conditions of an operation contain an unknown element about which *n* different hypotheses $A_1, A_2, \ldots, A_n$ constituting a complete system of events can be formulated. We somehow know their prior probabilities[13] $P(A_i)$ and we also know that according to hypothesis A_i some event *K* (for example, a hit of the target) has probability $P_{A_i}(K)$, $1 \le i \le n$. This is the probability of event *K* provided that hypothesis A_i is true.

If K occurred as the result of an experiment, the probabilities of the hypotheses A_i ought to be reappraised, and their new probabilities $P_K(A_i)$ determined. This is what the Bayes formula does.

Artillery firing begins by preliminary shots for specifying the location of the target. The unknown element can also be any other condition of firing (in particular, some peculiar feature of the gun **[?]**). Very often a few such shots are fired, and the problem consists then in calculating the new probabilities of the hypotheses on the basis of the results obtained. In all such cases the Bayes formula easily solves the problem at hand[14].

For the sake of brevity we denote

$$P(A_i) = P_i,\ P_{A_i}(K) = p_i,\ 1 \le i \le k\ \textbf{[}n\textbf{]}.$$

The Bayes formula is then written simpler:

$$P_K(A_i) = \frac{P_i p_i}{\sum_{r=1}^{n} P_r p_r},\ 1 \le i \le n.$$

Suppose that a volley of s trial shots were fired and result K appeared m times and failed $(s - m)$ times. Denote by K^* this result of the volley. We may assume that the results of the separate shots are mutually independent events. If hypothesis A_i is true the probability of result K is p_i and consequently the probability of the contrary event, of the failure of K, is $(1 - p_i)$. Then, by the multiplication rule for mutually independent events, the probability that the result K occurred after m definite shots is $p_i^m(1 - p_i)^{s-m}$.

Any m shots out of s can be selected and K can therefore occur in C_s^m incompatible ways (in the number of combinations of s elements taken m at a time). By the addition rule we therefore have

$$P_{A_i}(K^*) = C_s^m p_i^m (1 - p_i)^{s-m},\ 1 \le i \le n$$

and according to the Bayes formula

$$P_{K^*}(A_i) = \frac{P_i p_i^m (1 - p_i)^{s-m}}{\sum_{r=1}^{n} P_r p_r^m (1 - p_r)^{s-m}},\ 1 \le i \le n. \qquad (4.11)$$

This is the solution of the problem. Such problems occur not only in artillery but in other fields of human activities as well.

Example 1. Return to the beginning of this section. Required is the probability that the target is situated in interval a if two shots aimed at that interval were successful.

Denote a reiterated hit by K^*. By formula (4.11) we have

$$P_{K^*}(A) = \frac{P(a)[P_a(K)]^2}{P(a)[P_a(K)]^2 + P(b_1)[P_{b_1}(K)]^2 + \ldots}.$$

A simple calculation which we leave for the readers will convince them that the probability that the target is situated in interval a will heighten still more.

Example 2. An article manufactured at a certain factory is standard with probability 0.96. The articles are tested in a simplified way: a positive answer[1] is provided with probability 0.98 if the article is standard but only with probability 0.05 otherwise. Required is the probability that an article which stood two tests is standard.

Here, the complete system of hypotheses is composed of two contrary events: the article is, or is not standard. Their prior probabilities are $P_1 = 0.96$ and $P_2 = 0.04$. Under these hypotheses the probabilities that an article will stand the test are $p_1 = 0.98$ and $p_2 = 0.05$. After the two tests the probability of the first hypothesis, according to formula (4.11), will be

$$\frac{P_1 p_1^2}{P_1 p_1^2 + P_2 p_2^2} = \frac{0.96 \cdot 0.98^2}{0.96 \cdot 0.98^2 + 0.04 \cdot 0.05^2} \approx 0.9999.$$

In one case out of ten thousand we can be mistaken and consider a standard article substandard. Usually this result is good enough.

Example 3. After examining a patient three illnesses A_1, A_2 and A_3 were suspected and their probabilities under given conditions were

$P_1 = 1/2$, $P_2 = 1/6$ and $P_3 = 1/3$.

An additional analysis, which provides a positive answer with probabilities 0.1, 0.2 and 0.9 respectively, was prescribed and carried out 5 times. In four cases the result was positive and required are the probabilities of each illness after these analyses.

By the multiplication rule in case of illness A_1 the probabilities of the stated outcomes of the analyses are

$p_1 = C_5^4 0.1^4 \cdot 0.9$, $p_2 = C_5^4 0.2^4 \cdot 0.8$, $p_3 = C_5^4 0.9^4 \cdot 0.1$.

According to the Bayes formula we find that after the analyses the probabilities of those illnesses become respectively,

$$\frac{P_1 p_1,\ P_2 p_2,\ P_3 p_3}{P_1 p_1 + P_2 p_2 + P_3 p_3}.$$

When substituting the data, we have identical denominators

$1/2 \cdot 0.1^4 \cdot 0.9 + 1/6 \cdot 0.2^4 \cdot 0.8 + 1/3 \cdot 0.9^4 \cdot 0.1$

and numerators

$1/2 \cdot 0.1^4 \cdot 0.9$, $1/6 \cdot 0.2^4 \cdot 0.8$, $1/3 \cdot 0.9^4 \cdot 0.1$.

Calculation provides probabilities ca. 0.002, ca. 0.01 and ca. 0.988. The three events A_1, A_2 and A_3 constitute, as they had previously, a complete system of events so that the sum of the derived probabilities is unity, again as previously, which serves for checking the calculation.

Chapter 5. The Bernoulli Pattern

5.1. Examples. The length of about 75% of the fibres of cotton of a certain sort is shorter than 45 *mm* (short fibres or short) and about 25%, longer than, or equal to 45 *mm* (long fibres or long). Required is the probability that two out of three randomly selected specimens are short and one is long.

Denote by *A* the event of selecting a short fibre and by *B*, the contrary event. Then, evidently, P(*A*) = 3/4, P(*B*) = 1/4. Denote also by *AAB* the compound event consisting of two short first specimens and a long third specimen. The meaning of notation *BBA*, *ABA* etc is evident. Required is the probability of event *C*, of two short fibres and one long. It occurs in three possible ways,

AAB, *ABA* and *BBA*. (5.1)

Any two of them are mutually inconsistent and by the addition rule

$$P(C) = P(AAB) + P(ABA) + P(BAA).$$

The terms in the right side are identical since the selection of the specimens can be assumed mutually independent. According to the multiplication rule for mutually independent events the probability of each pattern (5.1) is a product of three factors two of which are P(*A*) = 3/4 and one is P(*B*) = 1/4 and thus is $(3/4)^2 \cdot 1/4 = 9/64$ and

$$P(C) = 3 \cdot 9/64 = 27/64,$$

which is our answer.

Example 2. Observations lasting many decades have established that out of a thousand births 515 newborn babies are boys[16] and 485 are girls. A family has 6 children and it is required to determine the probability that among them there are not more than 2 girls.

For the studied event to occur there should be 0, 1 or 2 girls; denote the respective probabilities of those events by P_0, P_1 and P_2. By the addition rule

$$P = P_0 + P_1 + P_2. \quad (5.2)$$

It is easiest to determine P_0. A male or female birth can be considered independent from the births of the other babies and by the multiplication rule the probability of all six male births is

$$P_0 = 0.515^6 \approx 0.018.$$

There are 6 ways for only one girl in the family: she can be the first, the second, …, the sixth child. Suppose that she was the fourth. According to the multiplication rule the probability of this case is equal to the product of six fractions, five of them equal to 0.515 and one to 0.485. It thus equals $0.515^5 \cdot 0.485$. All the other possible cases have the same probability and by the addition rule

$P_1 = 6 \cdot 0.515^5 \cdot 0.485 \approx 0.105.$

Now P_2. We note at once that there are many possibilities for the birth of two girls (for example, both the second and the fifth babies are girls, the other babies are boys). The probability of each such case is $0.515^4 \cdot 0.485^2$ which should be multiplied by the number of the possible cases. That number is $C_6^2 = 15$ and

$P_2 = 15 \cdot 0.515^4 \cdot 0.485^2 \approx 0.247.$

Finally

$P = P_0 + P_1 + P_2 \approx 0.018 + 0.105 + 0.247 = 0.370.$

Somewhat rarer than in four cases out of ten (with probability $P \approx 0.37$) the number of girls in such families will not be more than 1/3 of all the children (and not less than 2/3 of them will be boys).

5.2. The Bernoulli Formulas. In the previous sections we became acquainted with *repetitions of trials* with a certain event *A* possibly occurring in each. The word *trial* has many various meanings. When firing at a target, each shot is a trial; when the working life of light bulbs is ascertained, the test of each is a trial; when the structure of many newborn babies is studied with respect to sex, weight or height, the examination of each is a trial. In general, we will understand a trial as a realization of some conditions under which a studied event can happen.

Consider now one of the main patterns of probability theory which has many applications in various branches of science and is very important for that mathematical theory itself. This pattern consists in a sequence of mutually independent trials, i. e., such that the probability of one or another result in any of them does not depend on the results of previous or future trials. In addition, according to this pattern a certain event *A* can occur or not **[?]** in each trial with probability *p* independent from the number of the trial. The pattern came to be called after Jakob Bernoulli, an **[a most]** eminent Swiss mathematician of the end of the 17^{th} century.

We have already considered the Bernoulli pattern in our examples; suffice it to recall those of the previous section. Now, however, we are studying the following general problem whose particular cases were considered in all the examples of this chapter.

Problem. Under some conditions the probability that event *A* occurs in each trial is *p*. Required is the probability that in *n* mutually independent trials it will appear *k* times and fail $(n - k)$ times.

The event whose probability is sought can be broken down into a series of events. For obtaining one such event we ought to select arbitrarily some *k* trials and assume that that event *A* indeed took place in each of those trials and failed in the other $(n - k)$ trials. Each such event therefore requires the occurrence of *n* definite results, of *k* occurrences and $(n - k)$ failures of the event *A*.

By the multiplication rule the probability of each such event is

$p^k(1-p)^{n-k}$

and the number of them is equal to C_n^k, to the number of n elements taken k at a time. Apply the addition rule and the known formula

$$C_n^k = \frac{n(n-1)...(n-k+1)}{k!}$$

for determining the probability sought. It will be

$$P_n(k) = \frac{n(n-1)...(n-k+1)}{k!} p^k(1-p)^{n-k}. \qquad (5.3)$$

It is often expedient to express C_n^k in a somewhat different way. Multiply its numerator and nominator by

$(n-k)(n-k-1) \ldots 2 \cdot 1.$

Then

$$C_n^k = \frac{n!}{k!(n-k)!}$$

where by definition 0! = 1. We have now

$$P_n(k) = C_n^k p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}. \qquad (5.4)$$

Formulas (5.3) and (5.4) are usually named after Bernoulli[17]. For large values of n and k the determination of $P_n(k)$ is difficult since the factorials $n!$, $k!$, and $(n-k)!$ are very large and awkwardly calculable numbers. They are therefore determined with the aid of special tables of factorials and some approximation formulas.

Example. The probability that the expenditure of water in a certain factory will be normal (will not exceed a definite volume) is 3/4. Required is the probability that the expenditure will remain normal during the next 1, 2, …, 5, 6 days.

Denote by $P_6(k)$ the probability that during k out of the 6 days the expenditure will be normal and calculate it by formula (5.4) taking p = 3/4:

$P_6(6) = (3/4)^6$; $P_6(5) = 6(3/4)^5 \cdot 1/4$; $P_6(4) = C_6^4 (3/4)^4 \cdot (1/4)^2$;
$P_6(3) = C_6^3 (3/4)^3 \cdot (1/4)^3$; $P_6(2) = C_6^2 (3/4)^2 \cdot (1/4)^4$;
$P_6(1) = 6(3/4) \cdot (1/4)^5$.

Finally, $P_6(0) = (1/4)^6$ is the probability that the expenditure will be excessive all the six days. The denominator of all seven fractions is 4^6

= 4096 which we will certainly bear in mind when finally calculating them. The result is

$P_6(6) \approx 0.18$, $P_6(5) \approx 0.36$, $P_6(4) \approx 0.30$, $P_6(3) \approx 0.13$,
$P_6(2) \approx 0.03$, $P_6(1) \approx P_6(0) \approx 0$.

The most probable excessive expenditure takes place during one or two days in six whereas the probability of such expenditure during five or six days [$P_6(1)$ or $P_6(0)$] is practically zero.

5.3. The Most Probable Number of the Occurrences of an Event. The previous example shows that the probability of a normal expenditure of water during exactly k days increases with k, takes its maximal value and begins to decrease which is best seen on a diagram[18]. A still clearer picture is provided by a diagram showing the change of $P_n(k)$ with k when n becomes large.

It is sometimes necessary to know *the most probable* number of the occurrences of an event, to know the value of k for which $P_n(k)$ is maximal (certainly with given p and n). In all cases the Bernoulli formulas allow us to solve simply this problem which is what we now describe.

We begin by calculating $P_n(k + 1)/P_n(k)$. By formula (5.4)

$$P_n(k + 1) = \frac{n!}{(k+1)!(n-k-1)!} p^{k+1}(1-p)^{n-k-1} \qquad (5.5)$$

and formulas (5.3) and (5.5) provide

$$\frac{P_n(k+1)}{P_n(k)} = \frac{n!k!(n-k)!\,p^{k+1}(1-p)^{n-k-1}}{n!(k+1)!(n-k-1)!\,p^k(1-p)^{n-k}} = \frac{n-k}{k+1}\frac{p}{1-p}.$$

The probability $P_n(k + 1)$ will be higher, equal or lower than $P_n(k)$ depending on which of the three expressions

$$\frac{n-k}{k+1}\frac{p}{1-p} > 1, = 1 \text{ and } < 1 \qquad (5.6)$$

takes place. If, for example, we wish to know the values of k which satisfy the first inequality, we arrive at

$np - (1 - p) > k$, $k < np - (1 - p)$.

And, until the increasing k becomes equal to that difference, we will have $P_n(k + 1) > P_n(k)$. Probability $P_n(k)$ will heighten with the increase of k. Thus, for $p = 1/2$ and $n = 15$, $np - (1 - p) = 7$ and, until $k < 7$, $P_n(k + 1) > P_n(k)$. Just the same, by issuing from the two other relations (5.6) we establish that

$P_n(k + 1) = P_n(k)$ if $k = np - (1 - p)$ and
$P_n(k + 1) < P_n(k)$ if $k > np - (1 - p)$.

As soon as an increasing k oversteps the boundary $np - (1 - p)$, the probability $P_n(k)$ will begin to decrease until reaching $P_n(n)$. This conclusion first of all confirms that the behaviour of the magnitude $P_n(k)$ with an increasing k as manifested in the example above is a general law which takes place in all cases (at first $P_n(k)$ increases, then decreases if only p is not too near to 0 or 1).

In addition, this conclusion allows us to solve immediately our problem, the determination of the most probable value of k (which we denote by k_0). For this value $P_n(k_0 + 1) \leq P_n(k_0)$ and $k_0 \geq np - (1 - p)$. On the other hand, $P_n(k_0 - 1) \leq P_n(k_0)$ so that, similar to the above,

$$k_0 - 1 \leq np - (1 - p) \text{ or } k_0 \leq np - (1 - p) + 1 = np + p.$$

The most probable value k_0 of k thus ought to satisfy the inequalities

$$np - (1 - p) \leq k_0 \leq np + p. \qquad (5.7)$$

A simple subtraction shows that the length of the interval $[np - (1 - p), np + p]$ in which that k_0 should be contained is unity. Therefore, if one end of that interval [for example, $np - (1 - p)$] is not an integer, that interval will without fail include one and only one integer and k_0 will be determined uniquely. We ought to consider this case normal: since $p < 1$ and $np - (1 - p)$ is only an integer in exceptional instances in which inequalities (5.7) provide two values of k_0, $np - (1 - p)$ and $np + p$ differing from each other by a unity. These values will indeed be most probable. Their probabilities coincide and are higher than the probabilities of all other values of k.

Here is an example of such an exceptional case. Let $n = 15$, $p = 1/2$, then $np - (1 - p) = 7$, $np + p = 8$. The most probable values of the number k of the arrival of the studied event are 7 and 8. Their probabilities coincide and are approximately equal 0.196.

Example 1. Observations of many years in a certain region established that the probability of rain on 1 July is 4/17. Required is the most probable number of that event during the next 50 years.

Here

$$n = 50,\ p = 4/17,\ np - (1 - p) = 50 \cdot 4/17 - 3/17 = 11.$$

This is an integer, so are dealing with an exceptional case[19]. The most, and equally probable numbers of rainy days will be 11 and 12.

Example 2. Particles of a certain type are observed in a physical experiment. Under the same conditions 60 particles appear in the mean during a definite time interval and each with probability 0.7 has velocities exceeding v_0. Under other conditions during the same time interval only 50 particles were observed in the mean but the probability of their velocities exceeding v_0 was 0.8. Under which conditions was the most probable number of *rapid* particles?

First conditions: $n = 60$, $p = 0.7$, $np - (1 - p) = 41.7$, $k = 42$

Second conditions, respectively: 50, 0.8, 39.8 and 40.

The most probable number of *rapid* particles is somewhat larger in the first case.

Number n is often very large (in artillery firing, in mass production of articles etc) and np will be very large as well (if only probability p is not exceedingly low). The second terms, $(1 - p)$ and p, of magnitudes $np - (1 - p)$ and $np + p$, the end points of the interval within which the most probable number of the occurrences of the studied event is situated, are less than unity. Both those magnitudes, and consequently the most probable number of the occurrences of the studied event as well, are near to np.

Thus, if the probability of connecting two people by telephone less than in 15 *sec* is 0.74, we may assume $1000 \cdot 0.74$ as the most probable number of such connections out of a thousand calls arriving at a telephone exchange. This conclusion can be formulated more precisely:

Let k_o be the most probable number of the occurrences of the studied event in n mutually independent trials. Then k_o/n is the relative frequency of those occurrences. Inequalities (5.7) provide:

$$p - (1 - p)/n \leq k_o/n \leq p + p/n.$$

Suppose that with an invariable probability of the occurrence of that event in a single trial we ever more increase the number of trials (the most probable number k_o of those occurrences will also increase). The fractions $(1 - p)/n$ and p/n in the inequalities above will become ever smaller and they can therefore, when n is large, be neglected and both $p - (1 - p)/n$ and $p + p/n$ (and consequently k_o/n) will then equal p.

With a large number of mutually independent trials the most probable relative frequency of the studied event k_o/n becomes practically equal to its probability in a separate trial.

If for a certain measurement the probability of making an error contained between α and β is 0.84, then, given a large number of measurements, we may expect that most probably in about 84% of cases the error will indeed be contained between α and β. This certainty does not mean that, with a large number of observations, the probability of having exactly 84% of such errors will be high. On the contrary, this *maximal probability* itself will then be very low. Above, just before Example 1, even for $n = 15$ the maximal probability was only 0.196.

That probability is only maximal in the relative sense: the probability of having 84% errors of measurement contained between α and β is higher than that of having 83 or 86% of such errors. On the other hand, it is easy to understand that in case of long series of independent measurements the probability of one or another number of errors of a given magnitude cannot be really interesting. For example, with 200 measurements it is hardly expedient to calculate the probability that exactly 137 of them are measured with a given precision. It is practically indifferent whether that number is 137, 136 or 138 or even 140. On the contrary, it is undoubtedly interesting for practical reasons to know the probability that the number of measurements with errors contained within given boundaries is larger than 100, or between 100 and 125, or less than 50 etc.

How to express such probabilities? Suppose we wish to determine the probability that the number of measurements with a given precision k is contained between 100 and 120 (and including 120). More specifically, we wish to determine the probability of inequalities

$$100 < k \leq 120$$

so that k should be equal to one of the numbers 101, 102, …, 119, 120. By the addition rule that probability is

$$P(100 < k \leq 120) = P_{200}(101) + P_{200}(102) + \ldots + P_{200}(120).$$

Given such large numbers, direct calculation of 20 separate probabilities of the kind $P_n(k)$ according to formula (5.4) will be very difficult and is never attempted. There exist convenient tables and approximation formulas whose compilation/derivation is based on complicated methods of mathematical analysis which we will not discuss. However, simple reasoning about probabilities of the kind $P(100 < k \leq 120)$ can provide information which leads to exhausting solutions of problems at hand. We describe this topic in the next chapter.

Chapter 6. The Bernoulli Theorem

6.1. Its Content. A diagram in § 5.3 [not reproduced here] shows probabilities $P_{15}(k)$ as a function of k. It is seen that for intervals of the same length, $2 \le k < 5$ and $7 \le k < 10$, the sums of the corresponding probabilities essentially differ. In general, as we know, the probabilities $P_n(k)$ increase with k, pass their maximal values and decrease. It is therefore clear that out of two such intervals of the same length the sum of the probabilities will be higher for the interval situated nearer to the most probable value k_0.

Here, however, much more can be stated. For n trials the number k has $(n + 1)$ possible values, $0 \le k \le n$. Select the interval only containing a small part (a hundredth, say) of all such values with midpoint k_0. It turns out that for very large values of n that interval corresponds to an overwhelming probability whereas all the other values of k taken together have an insignificantly low probability. Although the length of the selected interval is trifling as compared with n [recall: $0 \le k \le n$], the sum of the corresponding probabilities will be considerably higher than the probability corresponding to all the other values of k.

All this practically means that

With a series of a large number n of mutually independent trials, we may expect with probability near to unity that the number of the occurrences of event A is very near to its most probable value and only differs from it by a negligible part of n.

This proposition known as the Bernoulli theorem[20] and discovered at the end of the 17th century is one of the most important laws of probability theory. Until the mid-19th century all its proofs required complicated mathematical means, but then the great Russian mathematician P. L. Chebyshev justified it very simply and briefly. We provide now his remarkable proof.

6.2. Its Proof. We already know that in case of a large number n of trials the most probable number k_0 of the occurrences of event A barely differs from np where, as always, p is the probability of the occurrence of A in a separate trial. It is therefore sufficient to prove that in case of a large number of mutually independent trials the number k of the occurrences of A will with an overwhelming probability very little differ from np, differ not more than by an arbitrarily small part of number n (for example, not more than by $0.01n$ or $0.001n$ or, in general, not more than by εn with ε being an arbitrarily small definite number). In other words, we ought to prove that in case of a sufficiently large n the probability

$$P(|k - np| > \varepsilon n) \qquad (6.1)$$

will become arbitrarily low.

For ensuring this fact note that by the addition rule the probability (6.1) equals the sum of probabilities $P_n(k)$ for all those values of n which are contained in either direction more than εn apart from np. Since the sum of all probabilities of a complete system of events is unity, the Bernoulli theorem means that the overwhelming part almost

equal to unity of that sum corresponds to the interval [– εn, εn] with midpoint np, and only its insignificant part is left for the regions beyond that interval. And so,

$$\mathrm{P}(|k - np| > \varepsilon n) = \sum_{|k\ -\ np| > \varepsilon n} \mathrm{P}_n(k). \qquad (6.2)$$

We turn now to Chebyshev's reasoning. In each term of the sum written just above

$$\frac{|k\ -\ np|}{\varepsilon n} > 1, \text{ and therefore } \left(\frac{k-np}{\varepsilon n}\right)^2 > 1$$

so that the sum will only increase if each $\mathrm{P}_n(k)$ is multiplied by the left side of the latter inequality. Therefore

$$\mathrm{P}(|k - np| > \varepsilon n) <$$
$$\sum_{|k\ -\ np| > \varepsilon n} \left(\frac{k-np}{\varepsilon n}\right)^2 \mathrm{P}_n(k) = \frac{1}{\varepsilon^2 n^2} \sum_{|k\ -\ np| > \varepsilon n} (k-np)^2 \mathrm{P}_n(k).$$

The appeared sum will increase still more if we add new terms so that k will change not only from 0 to $np - \varepsilon n$ and from $np + \varepsilon n$ to n, but over the whole interval [0, n]. Therefore, all the more

$$\mathrm{P}(|k - np| > \varepsilon n) < \frac{1}{\varepsilon^2 n^2} \sum_{k=0}^{n} (k-np)^2 \mathrm{P}_n(k). \qquad (6.3)$$

This sum favourably differs from all the previous sums in that it can by precisely calculated. The Chebyshev method indeed consists in replacing the sum (6.2) which is difficult to calculate by the sum (6.3).

Now the calculation itself. No matter how long it appears, the difficulties will only be technical and everyone knowing algebra will overcome them. At first we easily determine

$$\sum_{k=0}^{n} (k-np)^2 \mathrm{P}_n(k) = \sum_{k=0}^{n} k^2 \mathrm{P}_n(k) - 2np \sum_{k=0}^{n} k\mathrm{P}_n(k) + n^2 p^2 \sum_{k=0}^{n} \mathrm{P}_n(k). \quad (6.4)$$

The last of the three terms on the right side is the sum of the probabilities of a complete system of events and equals unity. In each of the other two terms the summands corresponding to $k = 0$ are zero, and we may begin the summing from $k = 1$.

1. Express $\mathrm{P}_n(k)$ according to formula (5.4):

$$\sum_{k=1}^{n} k\mathrm{P}_n(k) = \sum_{k=1}^{n} \frac{kn!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

Since $n! = n(n-1)!$ and $k! = k(k-1)!$, we get

$$\sum_{k=1}^{n} k\mathrm{P}_n(k) = np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)![(n-1)-(k-1))]!}\, p^{k-1}(1-p)^{(n-1)\,-\,(k-1)}.$$

Set $l = k - 1$ with l changing from 0 to $n - 1$ rather than from 1 to n as k did:

$$\sum_{k=1}^{n} k\mathrm{P}_n(k) = np\sum_{l=0}^{n-1} \frac{(n-1)!}{l![(n-1-l)]!}\, p^{l}(1-p)^{n-1-l)} = np\sum_{l=0}^{n-1} \mathrm{P}_{n-1}(l).$$

On the right side the sum of the probabilities of a complete system of events (of all possible occurrences of event A in $(n - 1)$ trials) is obviously unity. Thus,

$$\sum_{k=1}^{n} k\mathrm{P}_n(k) = np. \qquad (6.5)$$

2. For calculating the first term of (6.4) we first derive

$$\sum_{k=1}^{n} k(k-1)\mathrm{P}_n(k).$$

The summand corresponding to $k = 1$ is zero and we begin with $k = 2$. Note that $n! = n(n - 1)(n - 2)!$ and $k! = k(k - 1)(k - 2)!$. We easily determine, after setting similar to the above $m = k - 2$:

$$\sum_{k=2}^{n} k(k-1)\mathrm{P}_n(k) = \sum_{k=2}^{n} \frac{k(k-1)n!}{k![(n-k)]!}\, p^{k}(1-p)^{n-k} =$$

$$n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)![(n-2)-(k-2))]!}\, p^{k-2}(1-p)^{(n-2)-(k-2)} =$$

$$n(n-1)p^2 \sum_{m=0}^{n-2} \frac{(n-2)!}{m!(n-2-m)!}\, p^{m}(1-p)^{n-2-m} =$$

$$n(n-1)p^2 \sum_{m=0}^{n-2} P_{n-2}(m) = n(n-1)p^2. \qquad (6.6)$$

The last equality appeared since the sum of the terms $\mathrm{P}_{n-2}(m)$ is the sum of the probabilities of some complete system of events, of all the possible numbers of the occurrences of event A in $(n - 2)$ trials.

Finally, formulas (6.5) and (6.6) lead to

$$\sum_{k=1}^{n} k^2\mathrm{P}_n(k) = \sum_{k=1}^{n} k(k-1)\mathrm{P}_n(k) + \sum_{k=1}^{n} k\mathrm{P}_n(k) = n^2p^2 + np(1-p). \qquad (6.7)$$

And now we substitute the results (6.5) and (6,7) into relation (6.4), then insert the appearing extremely simple expression into inequality (6.3):

$$\sum_{k=0}^{n}(k-np)^2 \mathrm{P}_n(k) = n^2p^2 + np(1-p) - 2np\cdot np + n^2p^2 = np(1-p),$$

$$\mathrm{P}(|k-np| > \varepsilon n) < \frac{np(1-p)}{\varepsilon^2 n^2} = \frac{p(1-p)}{\varepsilon^2 n}. \qquad (6.8)$$

This new inequality provides everything we needed. Indeed, we may select an arbitrarily small ε but then leave it fixed. On the other hand, according to the meaning of our statement, the number of trials n can be as large as we wish so that the fraction $p(1-p)/\varepsilon^2 n$ becomes arbitrarily small: with an increasing n its denominator can become as large as desired whereas its numerator does not change.

Let $p = 0.75$, then $(1-p) = 0.25$, $p(1-p) = 0.1875 < 0.2$. Choose $\varepsilon = 0.01$, then inequality (6.8) provides

$$\mathrm{P}(|k - \frac{3}{4}n| > \frac{n}{100}) < \frac{0.2}{0.0001n} = \frac{2000}{n}.$$

For $n = 200{,}000$, $\mathrm{P}(|k - 150{,}000|) < 0.01$. This actually means that, for example, having a settled process ensuring that 75% of the manufactured articles are in the mean of sort I, from 148,000 to 152,000 of them out of 200,000 will with probability 0.99 (that is, almost certainly) posses this property.

Two remarks are necessary here. *First*, in practice, more precise estimates are applied although their justification is much more complicated.

Second, our rough estimate provided by inequality (6.8) becomes essentially more precise when p is very low, or, on the contrary, near to unity. Thus, in the example just above suppose that the probability of an article being of sort I is

$p = 0.95$, then $(1-p) = 0.05$, $p(1-p) < 0.05$.

With $\varepsilon = 0.005$ and $n = 200{,}000$,

$$\frac{p(1-p)}{\varepsilon^2 n} < \frac{0.05\cdot 1000{,}000}{25\cdot 200{,}000} = 0.01,$$

just as previously. But here $\varepsilon n = 1000$ rather than 2000 and, since $np = 190{,}000$, the number of articles having that property will be actually contained between 189,000 and 191,000.

With $p = 0.95$ the inequality (6.8) thus practically guarantees that the interval for the expected number of articles having the stated property is twice shorter than for the case in which $p = 0.75$:

$\mathrm{P}(|k - 190{,}000| > 1000) < 0.01$.

Problem. A quarter of workers in a certain branch of industry have secondary school education. Required is the most probable number of such workers in a random sample of 200,000 and an estimation of the probability that their actual number in the sample deviates from the most probable number not more than by 1.6%.

We issue from the fact that the probability of having that education is 1/4 for each worker in the sample; this is indeed the meaning of a random sample. Then, $n = 200{,}000$, $p = 1/4$, $k_0 = np = 50{,}000$, $p(1 - p) = 3/16$. We ought to calculate the probability that

$$|k - np| < 0.016np = 800$$

where k is the sought number of workers. Select ε so that $\varepsilon n = 800$, then $\varepsilon = 800/n = 0.004$. Formula (6.8) leads to

$$P(|k - 50{,}000| > 800) < \frac{3}{16 \cdot 0.000016 \cdot 200{,}000} \approx 0.06,$$

$$P(|k - 50{,}000| \leq 800) > 0.94.$$

The most probable number of such workers is 50,000 and the probability sought is higher than 0.94. Actually, it is much higher.

In concluding this chapter, we note that inequality (6.8) can be written as

$$P(|k/n - p| > \varepsilon) < \frac{p(1-p)}{\varepsilon^2 n}.$$

The fraction k/n is the relative frequency of the occurrence of event A in n trials. It follows that for any arbitrarily small but fixed ε the probability that the relative frequency deviates from the probability of event A more than by ε becomes arbitrarily low as n increases. This is similar to the stability of the relative frequencies discussed at the beginning of Chapter 1.

Part 2

Random Variables

Chapter 7. Random Variable and the Law of Distribution

7.1 Notion of Random Variable. Above, we have many times encountered magnitudes whose values were not constant but changed due to random influences. Thus, the number of boys out of a hundred newborns will not be the same for all hundreds. The length of fibres of a certain sort of cotton considerably varies not only with the region of growth but even if taken from the same bush and boll.

A few more examples. **1)** When firing from the same gun at the same target and setting the same distance **[and direction]** the shells nevertheless scatter. The distance between the gun and the point in which the shell falls varies, takes differing numerical values depending on unaccountable circumstances.

2) The velocity of a gas molecule does not remain constant, it changes owing to the collisions of that molecule with other molecules. Each molecule can collide with any other molecule and the variation of its velocity is purely random.

3) The yearly number of meteorites hitting the earth[21] is not constant but experiences considerable variations depending on many *random* circumstances.

4) The weight of grains of wheat grown on a certain plot is not definite but changes from grain to grain. It is impossible to allow for the influence of *all* the factors (quality of the plot, conditions of sunlight, availability of water etc) determining the growth of the grains and their weight *randomly* changes.

In spite of the heterogeneity of those examples all of them illustrate the same picture. In each of them we have a magnitude somehow characterizing the result of an operation (the counting of the meteorites, the measuring of the length of the fibres). However we try to uniform the conditions of their realization, each of those magnitudes can take various values depending on random differences in the eluding circumstances of these operations.

In the theory of probability each such magnitude is called a *random variable*. The examples above are already sufficient for convincing us in that their study is so important for applying the theory to most various branches of knowledge and practice.

To know a random variable certainly does not mean to know its numerical value[22]. Indeed, if, for example, a condenser had been working 5324 hours before perforation, the time of its uninterrupted work has thus taken a definite value and ceased to be a random variable. So what should we know about such a variable for obtaining all the possible information about it as about a *random variable*?

First of all, obviously, we ought to know all its possible numerical values. Thus, suppose that, as found out by tests, the working life of electronic tubes ranges from 2306 (minimal value) to 12,108 hours (maximal value). That magnitude can therefore take any value between those boundaries. In our third example above, the yearly number of meteorites can be any non-negative integer 0, 1, 2, …

However, the knowledge only of the list of possible values of a random variable is not yet sufficient for practically necessary estimations. Thus, if, in our second example, we consider a gas under

two differing temperatures, the possible numerical values of the velocity of its molecules will be the same and the set of these values will not provide any possibility of a comparative estimation of the temperature. At the same time, however, a difference of the temperatures indicates a very considerable difference in the state of the gas.

If we wish to estimate the temperature of a given amount of gas and only know the set of the possible values of the velocity of its molecules, we will naturally ask how often one or another velocity is observed. In other words, we obviously try to find out the *probabilities* of the different possible values of the studied random variable.

7.2. Notion of the Law of Distribution. We begin with a quite simple example, with firing at a target. When hitting a circle in its middle (region I), the shot gets 3 points; for hitting it elsewhere (region II), 2 points, and for missing (region III), 1 point[23].

Consider the number of these points as a random variable; its possible values are 1, 2, and 3. Denote their probabilities by p_1, p_2 and p_3, so that p_3, for example, corresponds to hitting region I. The possible values of the random variable under consideration are the same for all shots but their probabilities can essentially differ. Such differences obviously determine the differences between the skills of the shots. Thus, a very good shot possibly has probabilities $p_3 = 0.8$, $p_2 = 0.2$ and $p_1 = 0.0$; for an average and a quite poor shot, 0.3, 0.5 and 0.2 and 0.1, 0.3, 0.6 respectively.

If a shot fires 12 times, the possible numbers of hit-points occurring in each region are 0, 1, 2, …, 11, 12. By itself, this information does not yet allow us to judge his skill. On the contrary, we can only form an exhausting impression about it when finding out in addition the probabilities of the mentioned numbers.

Such is the invariable situation: knowing the probabilities of the various possible values of a random variable we will thus know how often to expect the occurrence of its more or less favourable values. This is apparently sufficient for judging the efficiency or quality of the pertinent operation. Practice shows that the knowledge of the probabilities of all the possible values of a studied random variable is indeed sufficient for solving any problem concerned with estimating its capacity as an indicator of the quality of the appropriate operation[24].

We conclude that for completely characterizing a random variable as such it is necessary and sufficient to know the list of all its possible values and the probabilities of each of them.

A random variable is thus expediently described by a table with two rows, values and probabilities. For the best shot (see example above), the number of points considered as a random variable can be represented by a table

values: 1, 2, 3; probabilities: 0, 0.2, 0.8. (I)

In general, for a random variable with possible values x_i and probabilities p_i the table will be

values: $x_1, x_2, \ldots, x_n$; probabilities: $p_1, p_2, \ldots, p_n$.

Such a table is called *the law of distribution* of the appropriate random variable. The knowledge of this law allows us to solve all problems connected with the variable at hand.

Problem. The number of points gained by a shot after one attempt has (I) as its law of distribution. Another shot has a different law of distribution:

values: 1, 2, 3; probabilities: 0.2, 0.5, 0.3. (II)

Required is the law of distribution of the sum of points achieved after a double shot. Such sums are clearly random variables, and we are asked to compile a table for our example. We should therefore consider all possible results of a combined firing of our shots. In the following table we entered the probabilities of each result calculated by the multiplication rule for independent events. The numbers of points gained by the shots are denoted, respectively, by ξ and η.

[The authors' table lists the 9 possible results with the corresponding ξ, η, $\xi + \eta$ and the probability of that sum.]

The table shows that the sum $\xi + \eta$ takes values 3, 4, 5 and 6. Value 2 is impossible since its probability is zero[25]. Now, value 3 is achieved in two ways and by the addition rule its probability is 0 + 0.04. The arrival of one of the following results […] is necessary and sufficient for $\xi + \eta = 4$. […]

We have thus compiled the table of the [law of] distribution for $\xi + \eta$ which completely solves the formulated problem:

values: 3, 4, 5, 6; probabilities: 0.04, 0.26, 0.46, 0.24. (III)

The sum of the probabilities is unity. Each law of distribution ought to possess this property since we deal here with the sum of the probabilities of all possible values of a random variable; that is, with the sum of the probabilities of some complete group of events. It is convenient to apply this property for checking the calculations made.

Chapter 8. The Mean Value

8.1. Determination of the Mean Value of a Random Variable.

Those two shots whom we have discussed just now, can achieve 3, 4, 5 or 6 points depending on random circumstances; the respective probabilities were shown in table (III). Now, suppose we ask: how many points are achieved by two shots after firing once each? We are unable to answer inasmuch as different attempts lead to differing results. However, for estimating the skill of our pair, we will certainly look at the *mean* result over a volley of firing rather than at one attempt whose result can be random[26]. So how many points in the mean are achieved after one attempt? Such a question is quite reasonable and can be definitely answered.

We reason in the following way. If the pair of shots fire a hundred times, the table of their law of distribution will show that about 4 times they achieve 3 points; about 26, 46 and 24 times they achieve 4, 5 and 6 points respectively. The sum of the points is

$$3\cdot 4 + 4\cdot 26 + 5\cdot 46 + 6\cdot 24 = 490.$$

Divide this number by 100 and get 4.9 points in the mean for an attempt, and this is our answer.

Instead of this method of calculation we could have divided each term by 100 even before summing them up. The simplest way of doing it is by dividing by 100 each second multiplier of each term and thus to return to the probabilities entered in table (III). The mean number of points achieved in each attempt made by the pair of shots will then be

$$3\cdot 0.04 + 4\cdot 0.26 + 5\cdot 0.46 + 6\cdot 0.24 = 4.9.$$

The terms here are obtained by multiplying each possible value of our random variable by its probability. In general, suppose that some random variable is defined by the table

values: $x_1, x_2, \ldots, x_k$; probabilities: $p_1, p_2, \ldots, p_k$.

To recall: if p_1 is the probability of the value x_1 of a random variable ξ, then, after n operations, x_1 will be observed about n_1 times, and $n_1/n = p_1$ so that $n_1 = np_1$. Just the same, x_2 will appear about n_2 times, $n_2 = np_2, \ldots$, and x_k will appear about $n_k = np_k$ times. And so, a series of n operations will contain in the mean

$n_1 = np_1$ such operations in which $\xi = x_1$,
$n_2 = np_2$ such operations in which $\xi = x_2, \ldots$,
$n_k = np_k$ such operations in which $\xi = x_k$.

The sum of the values of ξ in all n operations will be about

$$x_1n_1 + x_2n_2 + \ldots + x_kn_k = n(x_1p_1 + x_2p_2 + \ldots + x_kp_k)$$

and the mean value $\overline{\xi}$ of ξ corresponding to a single operation will be

$$\overline{\xi} = x_1p_1 + x_2p_2 + \ldots + x_kp_k.$$

We have thus arrived at the following important rule[27]:

For obtaining the mean value of a random variable each of its possible values should be multiplied by the corresponding probability and the calculated products summed up.

Of what benefit is the knowledge of the mean value of a random variable? To be more convincing, we begin by offering a few examples.

Example 1. Return once more to the two shots. The points they achieve are random variables whose laws of distribution we have derived in § 7.2. An attentive look at those laws is sufficient for deciding that the first shot is more skilful. Indeed, his probability of the best result (3 points) is considerably higher whereas the probabilities of the other (of the worst) results are higher for the second. Such a comparison does not however satisfy us since it is purely qualitative. Unlike the temperature, say, which directly estimates the heat of a physical body, here there is yet no measure, no such number which would have directly estimated the skill of those shots. And therefore it can always happen that a direct consideration will not provide any answer or that the answer will be arguable. Thus, instead of tables (I) and (II) having tables

values: 1, 2, 3; probabilities: 0.4, 0.1, 0.5 (I′)
values: 1, 2, 3; probabilities: 0.1, 0.6, 0.3 (II′)

we would have been hard put to decide at a glance which shot is better skilled. Indeed, the best result (3 points) is more probable for the first shot, but so is the worst result (1 point). On the contrary, 2 shots are more probable for the second shot.

And so, calculate now by the rule above the mean number of points for both shots:

$$1\cdot0.4 + 2\cdot0.1 + 3\cdot0.5 = 2.1;\ 1\cdot0.1 + 2\cdot0.6 + 3\cdot0.3 = 2.2.$$

In the mean, the second shot attained a bit more than the first and it certainly follows that the result of numerous firing will generally be somewhat more favourable for the second shot. We may now state for sure that the second shot is better skilled. The mean value of the number of points provided a convenient measure for easily and undoubtedly comparing the skills of the shots.

Example 2. When assembling a device, the most precise adjustment of its certain part can require 1, 2, 3, 4 or 5 attempts depending on luck. The number of attempts, ξ, is a random variable with those possible values. Suppose that their probabilities are given in the table:

values: 1, 2, 3, 4, 5; probabilities: 0.07, 0.16, 0.55, 0.21, 0.01.

If asked to supply as many parts as necessary for 20 devices[28], we will be unable to apply this table for estimating that number since it

only informs us that it varies from one case to another. However, if we determine the mean number $\overline{\xi}$ of attempts necessary for a device and multiply it by 20, we will obviously arrive at such an approximate number. We have

$$\overline{\xi} = 1\cdot 0.07 + 2\cdot 0.16 + 3\cdot 0.55 + 4\cdot 0.21 + 5\cdot 0.01 = 2.93,$$
$$20\,\overline{\xi} = 58.6 \approx 59.$$

It is reasonable to have an additional small reserve and prepare 60 – 65 parts.

In these examples, we needed some approximate estimate for a random variable. A glance at a table [of its law of distribution] will not provide such an estimate; it only informs us that the variable can take some values with some probabilities. However, the *mean value* of the random variable calculated by that table is already capable of furnishing such an estimate. It is indeed the value that the random variable will take in the mean in a more or less long series of operations. The mean value especially well characterizes a random variable when the operations are numerous or repeated many times over.

Problem 1. A series of trials is made with a constant probability p of the occurrence of some event A **[in each trial]** and the results of separate trials are independent from one another. Required is the mean frequency of the occurrence of A in n trials.

That frequency is a random variable with possible values 0, 1, 2, …, n, and the probability of some value k is, as we know (§ 6.2),

$$\mathrm{P}_n(k) = \frac{n!}{k!(n-k)!}\, p^k (1-p)^{n-k}.$$

The mean value sought is therefore

$$\sum_{k=0}^{n} k\mathrm{P}_n(k) = np$$

as calculated in that section. We have also convinced ourselves in that for any large n the most *probable* number of these occurrences is close to np.

In this case the most probable value of a random variable is near its mean value, but we ought to beware of believing that such closeness takes place for any random variable: these values can be very far apart. Thus, a random variable with the law of distribution

values: 0, 5, 10; probabilities: 0.7, 0.1, 0.2

has 0 as its most probable value whereas its mean value is 2.5.

Problem 2. Independent trials are made with probability 0.8 of the occurrence of some event A in each of them. Not more than 4 trials are carried out but, a second restriction, they only continue until the first appearance of A. Required is the mean number of those trials.

The number of trials can be 1, 2, 3 or 4 and we ought to determine their probabilities. In case of only one trial the event A should occur at once and the probability of this event is $p_1 = 0.8$. For the case of 2 trials it is necessary that the event only occurs at the second one after failing at the first one. By the multiplication rule for independent events

$$p_2 = (1 - 0.8)\cdot 0.8 = 0.16.$$

For the case of 3 trials similarly

$$p_3 = (1 - 0.8)^2\cdot 0.8 = 0.032$$

and for the last case, the event A should fail in the first three and either occur or fail in the fourth:

$$p_4 = (1 - 0.8)^3 = 0.008.$$

The number of trials considered as a random variable is determined by its law of distribution

values: 1, 2, 3, 4; probabilities: 0.8, 0.16, 0.032, 0.008.

The mean value of that number is

$$1\cdot 0.8 + 2\cdot 0.16 + 3\cdot 0.032 + 4\cdot 0.008 = 1.248.$$

Suppose that 100 such experiments should be carried out. We may then expect to carry out $1.248\cdot 100 \approx 125$ trials.

Such problems often occur in practice. For example, we test the strength of yarn. It is of top quality if it does not tear even once under a specified load during tests of not more than four specimens of standard length taken from the same skein or boll.

Problem 3. A side of a square plot as shown on an air survey photo is measured with possible errors[29] 0, ± 10, ± 20, ± 30 *m* having probabilities 0.42, 0.16, 0.08, 0.05. Required is the mean area of the plot as determined by these measurements.

The length of the side is a random variable with law of distribution

values: 320, 330, 340, 350, 360, 370, 280 *m*;
probabilities: 0.05, 0.08, 0.16, 0.42, 0.16, 0.08, 0.05. (I)

We can at once derive the mean value of that length, but in this case it is not even necessary: the same errors in each direction are equally probable and this symmetry leads to mean value 350 *m*. In more detail: the mean value includes terms

$350\cdot 0.42$;
$(340 + 360)\cdot 0.16 = [(350 - 10) + (350 + 10)]\cdot 0.16 = 2\cdot 350\cdot 0.16$;
$(330 + 370)\cdot 0.08 = 2\cdot 350\ 0.08$; $(320 + 380)\cdot 0.05 = 2\cdot 350\cdot 0.05$

and it therefore equals

$350(0.42 + 2\cdot0.16 + 2\cdot0.08 + 2\cdot0.05) = 350.$

We may surmise that the mean value of the area of the plot is $350^2 = 122{,}500\ m^2$. This result would have been correct had the mean value of the square of a random variable been equal to the square of its mean value. However, this premise is false. In our example, the possible values of the area of the square are

$320^2, 330^2, 340^2, 350^2, 360^2, 370^2, 380^2.$

Which is the true value? It depends on which of the seven cases represented in table (I) will take place. The probabilities of the seven possible values are therefore the same as shown in table (I). It follows that the law of distribution of the area is

values: as stated just above;
probabilities: 0.05, 0.08, 0.16, 0.42, 0.16, 0.08, 0.05

The mean value of the area is

$320^2\cdot0.05 + 330^2\cdot0.08 + 340^2\cdot0.16 + 350^2\cdot0.42 +$
$360^2\cdot0.16 + 370^2\cdot0.08 + 380^2\cdot0.05.$

Here also, symmetry, as it often occurs, simplifies calculation and it is worthwhile to show how exactly this simplification is achieved. We have

$350^2\cdot0.42 + (340^2 + 360^2)\cdot0.16 +$
$(330^2 + 370^2)\cdot0.08 + (320^2 + 380^2)\cdot0.05.$

Now,

$340^2 + 360^2 = (350 - 10)^2 + (350 + 10)^2,$
$330^2 + 370^2 = (350 - 20)^2 + (350 + 20)^2,$
$320^2 + 380^2 = (350 - 30)^2 + (350 + 30)^2,$

so that the sum above is

$350^2\cdot[0.42 + 2\cdot0.16 + 2\cdot0.08 + 2\cdot0.05] +$
$2\cdot10^2\cdot0.16 + 2\cdot20^2\cdot0.08 + 2\cdot30^2\cdot0.05 =$
$350^2 + 2(16 + 8 + 45) = 122{,}686\ (m^2).$

All this can be calculated mentally **[?]**. The mean value of the area of the square is somewhat (in this case, imperceptibly) larger than the square of the mean value of its side, $350^2 = 122{,}500\ (m^2)$. It is not difficult to show that such is the general rule: *the mean value of a square of any random variable is always*[30] *larger than the square of its mean value.*

Indeed, suppose we have a random variable ξ with a perfectly arbitrary law of distribution

values: $x_1, x_2, \ldots, x_k$; probabilities: $p_1, p_2, \ldots, p_k$.

The law of distribution of its square will be

values: $x_1^2, x_2^2, ..., x_k^2$; probabilities: the same as above.

Then

$$\bar{\xi} = x_1 p_1 + x_2 p_2 + ... + x_k p_k, \quad \bar{\xi}^2 = x_1^2 p_1 + x_2^2 p_2 + ... + x_k^2 p_k,$$
$$\bar{\xi}^2 - (\bar{\xi})^2 = \bar{\xi}^2 - 2(\bar{\xi})^2 + (\bar{\xi})^2.$$

Since the sum of the probabilities is unity, the three terms in the right side of the last equality are

$$\bar{\xi}^2 = \sum_{i=1}^{k} x_i^2 p_i, \quad 2(\bar{\xi})^2 = 2(\bar{\xi})(\bar{\xi}) = 2\bar{\xi}\sum_{i=1}^{k} x_i p_i = \sum_{i=1}^{k} 2\bar{\xi} x_i p_i,$$
$$(\bar{\xi})^2 = (\bar{\xi})^2 \sum_{i=1}^{k} p_i = \sum_{i=1}^{k} (\bar{\xi})^2 p_i.$$

Therefore

$$\bar{\xi}^2 - (\bar{\xi})^2 = \sum_{i=1}^{k} [x_i^2 - 2\bar{\xi} x_i + (\bar{\xi})^2] p_i = \sum_{i=1}^{k} (x_i - \bar{\xi})^2 p_i.$$

All the terms of the last sum are non-negative, therefore

$\bar{\xi}^2 - (\bar{\xi})^2 \geq 0$, QED.

Chapter 9. Mean Values of Sums and Products

9.1. A Theorem on the Mean Value of Sums. Very often it is necessary to calculate the mean value of a sum of two (and not rarely of a larger number of) random variables with known mean values. Suppose for example that two factories manufacture the same articles and that their daily produce is, in the mean, 120 and 180 of them respectively. Can we now establish the mean value of their combined daily produce? Or is the data insufficient and we ought to know in addition something else (for example, the pertinent laws of distribution)?

It is very important that the knowledge of the mean values of the summands is always sufficient for calculating the mean value of their sum. And that the latter is expressed through the former in the easiest possible way: *the mean value of a sum always equals the sum of the mean values of the summands*. Thus, if ξ and η are perfectly arbitrary random variables,

$$\overline{\xi+\eta} = \overline{\xi} + \overline{\eta}.$$

In the example above, $\overline{\xi} = 120$, $\overline{\eta} = 180$, $\overline{\xi+\eta} = \overline{\xi} + \overline{\eta} = 300$.

To prove this rule in the general case, suppose that the laws of distribution of those random variables are

values: $x_1, x_2, \ldots, x_k$; probabilities: $p_1, p_2, \ldots, p_k$; (I)
values: $y_1, y_2, \ldots, y_l$; probabilities: $q_1, q_2, \ldots, q_l$. (II)

The possible values of $\xi + \eta$ are all the sums of the kind of $x_i + y_j$, $1 \le i \le k$, $1 \le j \le l$. The probability of that sum, p_{ij}, is unknown. It is the probability of a joint event $\xi = x_i$ and $\eta = y_j$. Had these two events been independent, then, obviously, by the multiplication rule, we would have had

$$p_{ij} = p_i q_j, \tag{9.1}$$

but we will not at all assume that condition.

And so, equality (9.1) will not generally take place and we ought to take into account that the knowledge of the laws of distribution (I) and (II) does not in general allow us to conclude anything about the probability p_{ij}. According to the general rule, the mean value of the sum $\xi + \eta$ equals the sum of the products of all its possible values by their probabilities:

$$\overline{\xi+\eta} = \sum_{i=1}^{k}\sum_{j=1}^{l}(x_i + y_j)\,p_{ij} = \sum_{i=1}^{k} x_i\Big[\sum_{j=1}^{l} p_{ij}\Big] + \sum_{j=1}^{l} y_j\Big[\sum_{i=1}^{k} p_{ij}\Big]. \tag{9.2}$$

Consider attentively the fist sum of p_{ij}. It is the sum of the probabilities of all possible events $\xi = x_i$ and $\eta = y_j$ with i being the same in all the terms of that sum and j ranging over all of its possible values from 1 to l inclusive. Since the events $\eta = y_j$ with differing j are

obviously incompatible, that sum is by the addition rule the probability of one *out of the l events*, $\xi = x_i$ and $\eta = y_j$ where $j = 1$ or 2 or … or l.

However, to say that one such event had occurred is the same as saying that $\xi = x_i$. Indeed, if one such event did occur, then, clearly, the event $\xi = x_i$ had also appeared. Inversely, if event $\xi = x_i$ had occurred, than, since η ought to take one of its values $y_1, y_2, \ldots, y_l$, one of the events $\xi = x_i$ and $\eta = y_j$ ($j = 1$ or 2 or … or l) also happened[31].

The sum of p_{ij} with a constant i, being the probability of the occurrence of one of the events just mentioned, is simply equal to the probability of $\xi = x_i$; that is, to that very sum, to p_i. Just the same, we certainly convince ourselves in that the other sum of p_{ij} (with constant j) equals q_j. Setting these two expressions into the equality (9.2), we find that

$$\overline{\xi+\eta} = \sum_{i=1}^{k} x_i p_i + \sum_{j=1}^{l} y_i q_j = \overline{\xi} + \overline{\eta}, \text{ QED.}$$

We have proved this theorem for two summands, but it is immediately extended to three or more summands since

$$\overline{\xi+\eta+\zeta} = \overline{\xi+\eta} + \overline{\zeta} = \overline{\xi} + \overline{\eta} + \overline{\zeta} \text{ etc.}$$

Example. In a certain factory a manufactured article is selected from each of the n lathes. Determine the mean number of substandard articles if the probabilities of their production are respectively p_1, $p_2, \ldots, p_n$. The number of rejects per one article is a random variable with only two possible values, 1 and 0 whose probabilities are p_1 and $(1 - p_1)$, p_2 and $(1 - p_2)$ etc. The mean number of substandard articles selected from the first lathe is

$$1p_1 + 0(1 - p_1) = p_1.$$

The same magnitudes for the other lathes are $p_2, \ldots, p_n$ and the mean value of the total number of substandard articles is $p_1 + p_2 + \ldots + p_n$. In particular, if these probabilities coincide, that mean number will be pn.

We have already determined this result (6.5), but it is interesting to compare the awkward previous calculations with this simplest reasoning which did not require any calculations. Moreover, in addition to simplicity, we have gained generality. Previously, we assumed that the results of the separate trials were *mutually independent* and our conclusion was only valid under this condition. Now, however, we manage without it since the addition rule for mean values takes place for any random variables without any restrictions. And if p is constant, be there any dependence between the lathes and the articles, the mean number of the rejects will always remain without change, np.

9.2. A Theorem on the Mean Value of Products. The problem considered in § 9.1 often has to be also studied for the products of random variables. Suppose that ξ and η have, as previously, laws of

distribution (I) and (II). The product $\xi\eta$ is a random variable with possible values $x_i y_j$, $1 \le i \le k$ and $1 \le j \le l$, and probabilities p_{ij}. The problem consists now in formulating a rule allowing us to express the mean value $\overline{\xi\eta}$ of $\xi\eta$ through the mean values of ξ and η. A general solution of this problem is however impossible since the mean value sought is not uniquely determined by the mean values $\overline{\xi}$ and $\overline{\eta}$: various values of $\overline{\xi\eta}$ are possible for the same values of $\overline{\xi}$ and $\overline{\eta}$ and a general formula expressing the former through the latter is impossible.

Nevertheless, there exists an important exception and, moreover, the derived connection is then extremely simple. We will call the random variables ξ and η *independent* if for any i and j the events $\xi = x_i$ and $\eta = y_j$ are independent, if some definite value taken by one of the random variables does not influence the law of distribution of the other variable.

And so, if ξ and η are independent in the defined sense, then, by the multiplication rule for independent events,

$p_{ij} = p_i q_j$, $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, l$.

Therefore

$$\overline{\xi\eta} = \sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j p_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j p_i q_j = \sum_{i=1}^{k} x_i p_i \sum_{j=1}^{l} y_j q_j = \overline{\xi}\,\overline{\eta}.$$

The mean value of the product $\xi\eta$ *of independent random variables* ξ *and* η *equals the product of the mean values of its factors, of* ξ *and* η.

Just like it was in the previous case of addition, this rule derived for the product of two random variables immediately extends to the product of any number of factors. It is only necessary for those factors to be mutually independent so that the knowledge of definite values of some of those variables does not influence the laws of distribution of the other variables.

In case of dependent variables ξ and η the mean value of their product $\xi\eta$ can be unequal to the product of the mean values of ξ and η. Suppose for example that the law of distribution of ξ is

values: – 1, 1; probabilities: 0.5, 0.5

and that the distribution of another random variable $\eta = \xi$ is the same; the mean values of both these variables are zero, but $\xi\eta = \xi^2$ always equals 1, therefore $\overline{\xi\eta} = 1$. If, however, $\eta = -\xi$, its distribution remains as it was previously, but the product $\xi\eta$ always equals – 1 and $\overline{\xi\eta} = -1$.

Example. An electric current whose strength I depends on random circumstances flows along a conductor whose resistance R also depends on randomness. The mean value of the resistance is 25 *ohms* and the mean value of the current's strength is 6 *amp*. Required is the mean value of the drop of the voltage E.

According to the Ohm law, E equals the product RI. We have $\bar{R} = 25$, $\bar{I} = 6$. Assuming that these magnitudes are independent, we find that

$$\bar{E} = \overline{RI} = 25 \cdot 6 = 150 \text{ volts.}$$

Chapter 10. Scatter and Mean Deviations

10.1. The Mean Value Is Insufficient for Characterizing a Random Variable. Time and time again we have seen that the mean value of a random variable provides a rough guide for imagining it which is sufficient in many practical instances. Thus, for comparing the skill of two competing shots suffice it to know the mean values of their gained points. For comparing the efficiency of two differing systems of counting cosmic particles it is quite sufficient to know the number of those possibly skipped by each system etc. In each such case we considerably benefit by describing a random variable by a single number, by its mean value, rather than defining it by a complicated law of distribution. It appears then as though we are dealing with a positively known magnitude with a completely definite value.

Much oftener, however, we encounter a situation in which the mean value of a random variable does not determine its most practically important features. A more detailed acquaintance with its law of distribution is then required.

A typical case in point is the study of the distribution of the errors of measurement. Let ξ be the magnitude of an error, of a deviation of the obtained value of the measured magnitude from its mean value. If systematic errors are absent, the mean value of the error, $\overline{\xi}$, is zero. How then are the errors scattered? How often will errors of some magnitude occur? Only knowing that $\overline{\xi} = 0$, we have no answer to any of these questions. Often it is only known that both positive and negative errors are possible and that their probabilities approximately coincide. We do not know, however, the most important feature: are most results of measurement located near the true value of the measured magnitudes[32], so that we may reckon that each result is highly reliable, or are they mostly scattered over large intervals in each direction from that value. Both possibilities are encountered in practice.

Two observers measuring a certain magnitude with the same mean error $\overline{\xi}$ can produce results of differing degrees of precision. It can occur that the measurements of one of them systematically scatter more extensively which means that the absolute values of the errors of his measurements can be larger in the mean and that his results will deviate farther from the measured magnitude than the results of the other observer.

Another example. Two sorts of wheat are tested for crop capacity. Depending on random circumstances (quantity of rainfall, distribution of fertilizers, solar radiation etc) the yield per square meter is subject to considerable fluctuations and is a random variable. Suppose that under the same conditions the mean yield is the same in both cases, 240 g/m^2. Can we judge the quality of the sorts only by this mean yield? Apparently not since most practically useful is that sort whose yield is less exposed to random influences of meteorological and other factors, whose yield scatters less. And so, the possible fluctuation of the yield is not less important than its mean value.

10.2. Various Methods of Measuring the Scatter of a Random Variable. The examples above as well as **[possible]** similar illustrations convincingly indicate that in many cases the knowledge of the mean values of random variables is just insufficient for describing their most interesting features. Those features remain unknown, and we ought therefore to have their entire tables of distribution before our eyes which is almost always complicated and inconvenient. We can also try to describe the random variables by one or two similar additional numbers so that the joined small set of **[two or three]** numbers will provide a practically sufficient characteristic of their most essential features. Let us see how we can realize the latter possibility.

The described examples show that in many cases it is practically most important to know the possible deviations[33] of the actual values of a given random variable from its mean value, to know the degree of its scattering. Are those values for the most part closely grouped around the mean value (and therefore tightly grouped themselves) or, on the contrary, do most of them very markedly deviate from that value (with some of them necessarily considerably differing from each other)?

The rough pattern below helps to imagine clearly the difference just described. Consider two random variables with laws of distribution respectively

values: – 0.01, 0.01; probabilities: 0.5, 0.5 (I)
values: – 100, 100; probabilities: 0.5, 0.5 (II)

Both have zero mean values; however, the first always takes values very near to zero (and to each other) whereas the second can only take values sharply differing from zero (and from each other). For the former, the knowledge of its mean value also provides rough information about its actual possible values. However, the mean value of the latter is very considerably apart from such possible values and furnishes no idea about them. Those possible values are much more *scattered* than in the first case.

Our problem thus consists of finding a number which would give us a reasonable *measure of scattering* of a random variable and at least roughly indicate to us how large the expected deviations are from its mean value. The deviation of random variable ξ from its mean value $\overline{\xi}$, $\xi - \overline{\xi}$, is itself a random variable as well as $|\xi - \overline{\xi}|$ which characterizes that deviation regardless of its sign. And we wish to have a number which will roughly characterize that random deviation $\xi - \overline{\xi}$ and tell us how large, approximately, can it be. This question can be solved in many ways; most common are the following three.

10.2.1. *The mean deviation*. It is most natural to adopt the mean value of $|\xi - \overline{\xi}|$ as a rough value of that very random variable. This mean value is called the *mean deviation* of ξ. If ξ has the law of distribution

values: $x_1, x_2, \ldots, x_k$; probabilities: $p_1, p_2, \ldots, p_k$

the law for $|\xi - \overline{\xi}|$ will be

values: $|x_1 - \overline{\xi}|, |x_2 - \overline{\xi}|, \ldots, |x_k - \overline{\xi}|$; probabilities: $p_1, p_2, \ldots, p_k$.

Here, $\overline{\xi} = x_1p_1 + x_2p_2 + \ldots + x_kp_k$. We thus obtain the mean deviation M_ξ of ξ

$$M_\xi = \sum_{i=1}^{k} |x_i - \overline{\xi}| \, p_i$$

with $\overline{\xi}$ as written just above.

For variables with laws of distribution (I) and (II) we have $\overline{\xi} = 0$ and $M_\xi = 0.01$ and 100 respectively. However, these examples are trivial since the pertinent absolute deviations can only take one value and thus in both cases the essence of a random variable is forfeited.

Calculate now the mean deviation for the random variables with laws of distribution (I′) and (II′) in § 8.1. We saw there that the mean values of those variables were 2.1 and 2.2, very near to each other. The mean deviations for those variables are

$0.4|1 - 2.1| + 0.1|2 - 2.1| + 0.5|3 - 2.1| = 0.9$
$0.1|1 - 2.2| + 0.6|2 - 2.2| + 0.3|3 - 2.2| = 0.48$

For the second variable the mean deviation is almost twice less. Actually this obviously means that, although in the mean the shots gained approximately the same number of points, and in this sense can be thought equally skilled, the hit-points of the second shot are *uniform* to a much greater extent, are much less scattered. The first shot, while achieving the same number of points, fires irregularly, and his results are often both much better and much worse than his mean results.

10.2.2. *The mean square deviation*. It is indeed natural but also very inconvenient to measure the rough magnitude of a deviation by its mean value since calculations and estimations are often complicated and sometimes simply impossible. Usually another measure of deviations is introduced. Just as the deviation $\xi - \overline{\xi}$ of the random variable ξ from its mean value $\overline{\xi}$, the square $(\xi - \overline{\xi})^2$ of this deviation is a random variable. In our previous notation, its law of distribution is

values: $(x_1 - \overline{\xi})^2, (x_2 - \overline{\xi})^2, \ldots, (x_k - \overline{\xi})^2$; probabilities: $p_1, p_2, \ldots, p_k$

and the mean value of this square is

$$\sum_{i=1}^{k} (x_i - \overline{\xi})^2 \, p_i.$$

It provides an idea of the approximate value of the *square* of the deviation $\xi - \overline{\xi}$. Extracting a square root of this sum

$$Q_\xi = \sqrt{\sum_{i=1}^{k} (x_i - \overline{\xi})^2 p_i}$$

we obtain a measure which is capable of characterizing the approximate magnitude of the deviation itself, the *mean square deviation* of random variable ξ. Its square, Q_ξ^2 **[also displayed above]**, is the *variance* of that variable[34]. This new measure of the deviation is certainly somewhat more artificial than the mean deviation introduced above. Here, we follow a roundabout path: first, we deduce an approximate value of the *square*, and only after that, by extracting the square root, return to the deviation itself. On the other hand, as shown in the next section, the application of the mean square deviation Q_ξ considerably simplifies calculations. It is this circumstance that compels statisticians to apply mainly this measure.

Example. For the random variables defined by their laws of distribution (I´) and (II´) of § 8.1 we have respectively

$$Q_\xi^2 = 0.4(1 - 2.1)^2 + 0.1(2 - 2.1)^2 + 0.5(3 - 2.1)^2 = 0.89$$
$$Q_\xi^2 = 0.1(1 - 2.2)^2 + 0.6(2 - 2.2)^2 + 0.3(3 - 2.2)^2 = 0.36$$

The square roots of these magnitudes, i. e., the mean square deviations, are ca. 0.94 and 0.6. For the same random variables we have derived the mean deviations 0.9 and 0.48. Both measures are considerably larger for the first random variable and in each case we conclude that that variable is more scattered than the second.

Again, in each case the mean square deviation is larger than the mean deviation and it is easy to understand that the same should happen for any random variable. Indeed, the variance Q_ξ^2 being the mean value of the square of $|\xi - \overline{\xi}|$ cannot be less than the square of the mean value M_ξ of $|\xi - \overline{\xi}|$, see end of § 8.1, and $Q_\xi \geq M_\xi$ follows from $Q_\xi^2 \geq M_\xi^2$.

10.2.3. *Probable deviation*. Another method of characterizing scattering is often applied, especially in military operations. We describe it in terms of an example.

Suppose that an artillery gun fires in a certain direction with shots ranging over distance ξ. Now, this is a random variable whose mean value indicates *the centre of hit-points* with shells falling around it[35]. The deviation $\xi - \overline{\xi}$ of the studied random variable (of the range) from its mean value is at the same time the deviation of a hit-point from the centre of such points. Any estimate of $|\xi - \overline{\xi}|$ therefore measures the scatter of shells as well and is the most important indication of the quality of firing.

From the centre of hit-points mark a very small segment α in both directions along the line of firing. Only a small fraction of shells will

fall within the interval [– α, α]. In other words, for small values of α the probability of $|\xi - \overline{\xi}| < \alpha$ is very low. Lengthen now that interval by increasing the arbitrary α and the probability of a shell falling within it will heighten. If α is very large, practically all the shells will fall within the thus lengthened interval. Therefore, the probability of the inequality $|\xi - \overline{\xi}| < \alpha$ heightens from 0 to 1. At first, the probability of $|\xi - \overline{\xi}| > \alpha$, of the shell falling beyond the interval, will be higher, then, with a larger value of α, it will become lower. So there ought to exist some value α_0 of α for which the probabilities of a shell falling either within, or beyond the corresponding interval will coincide. Both inequalities

$$|\xi - \overline{\xi}| < \alpha_0 \text{ and } |\xi - \overline{\xi}| > \alpha_0$$

are then equally probable and their common probability is therefore 1/2. Here, we neglect the insignificantly low probability of the exact *equality* $|\xi - \overline{\xi}| = \alpha_0$.

This α_0 is unique. Its magnitude depends on the quality of the artillery guns. It is easily seen that the value of α_0, just as the mean or the mean square deviation, can serve as a measure of the scattering of the shells. Indeed, if α_0 is very small, a half of the shells fall within a very small interval which testifies to a comparatively insignificant scatter. On the contrary, a large α_0 shows that a half of the shells still falls beyond the corresponding **[long]** interval. This obviously indicates that the scatter of the shells is considerable.

That number, α_0, is usually called *the probable deviation* of ξ. The absolute deviation $|\xi - \overline{\xi}|$ can with the same probability be either larger or smaller than it. That deviation denoted by E_ξ is not more convenient for calculations than the mean deviation M_ξ and much less convenient than the mean square deviation Q_ξ but nevertheless it is indeed adopted in artillery for estimating all deviations. Below, we show why this practice usually does not lead to any difficulties.

10.3. Theorems on the Mean Square Deviation. Let us show that those deviations indeed possess special properties compelling us to prefer them to any other pertinent characteristics. The following problem has basic importance for applications.

Suppose that independent random variables $\xi_1, \xi_2, \ldots, \xi_n$ have mean square deviations $q_1, q_2, \ldots, q_n$. Denote

$$\xi_1 + \xi_2 + \ldots + \xi_n = S_n$$

and ask ourselves how to determine the mean square deviation Q of S_n.

In accordance with the addition rule for mean values

$$\overline{S}_n = \overline{\xi}_1 + \overline{\xi}_2 + \ldots + \overline{\xi}_n$$

so that

$$S_n - \overline{S}_n = (\xi_1 - \overline{\xi}_1) + (\xi_2 - \overline{\xi}_2) + \ldots + (\xi_n - \overline{\xi}_n),$$

$$(S_n - \overline{S}_n)^2 =$$

$$[\sum_{i=1}^{n}(\xi_i - \overline{\xi}_i)]^2 = \sum_{i=1}^{n}(\xi_i - \overline{\xi}_i)^2 + \sum_{i=1}^{n}\sum_{k=1}^{n}(\xi_i - \overline{\xi}_i)(\xi_k - \overline{\xi}_k),\ i \neq k. \quad (10.1)$$

Note that m. v. $(S_n - \overline{S}_n)^2 = Q^2$, m. v. $(\xi_i - \overline{\xi}_i)^2 = q_i^2$, $i = 1, 2, \ldots, n$ where m. v. is our [O. S.] notation for *mean value of*.

By the addition rule for mean values we have ($i \neq k$)

$$Q^2 = \sum_{i=1}^{n} q_i^2 + \sum_{i=1}^{n}\sum_{k=1}^{n} \text{m.v.}[(\xi_i - \overline{\xi}_i)(\xi_k - \overline{\xi}_k)]. \quad (10.2)$$

However, we assumed that ξ_i and ξ_k, again for $i \neq k$, are independent and by the multiplication rule for independent magnitudes we have

$$\text{m.v.}\,[(\xi_i - \overline{\xi}_i)(\xi_k - \overline{\xi}_k)] = \text{m.v.}\,(\xi_i - \overline{\xi}_i)\,\text{m.v.}\,(\xi_k - \overline{\xi}_k).$$

Both factors on the right side disappear since, for example, the first equals $\overline{\xi}_i - \overline{\xi}_i$ and (10.2) becomes

$$Q^2 = \sum_{i=1}^{n} q_i^2.$$

The variance of the sum of independent random variables equals the sum of their variances. One more very important rule for variances of independent random variables is thus added to the addition rule for mean values.

For the mean square deviations we immediately obtain

Q = square root of the right side of the previous formula. (10.3)

This possibility of simply expressing the mean square deviation of a sum through the mean square deviations of its terms provided these are independent is indeed one of the most important advantages of the mean square deviation over mean, probable and other kinds of deviations.

Example 1. Suppose that in a certain factory each manufactured article can be substandard with probability p independently from the other articles. The mean number of rejects out of n manufactured articles is np (Problem 1 in § 8.1). For roughly estimating how largely the actual number of substandard articles can deviate from this mean value we will find the mean square deviation of the number of those rejected from np. The easiest way to calculate it is by applying formula (10.3).

Indeed, we can consider the number of substandard articles as the sum of the numbers of such articles appearing out of each manufactured. We have acted in this way when discussing a similar example in § 9.1. And since we assume that these numbers are independent random variables, we may apply the addition rule for variances and calculate the mean square deviation Q of the total number of rejects by formula (10.3). The magnitudes $q_1, q_2, \ldots, q_n$ will then denote the mean square deviations of the number of substandard articles per each article.

The number of rejects ξ_i appearing when manufacturing article i is determined by table

value: 1, 0; probabilities p, $1 - p$,

so we have $\overline{\xi}_i = p$ and

$$q_i^2 = \text{m.v. } (\xi_i - \overline{\xi}_i)^2 = (1-p)^2 p + p^2(1-p) = p(1-p),$$

$$Q = \sqrt{\sum_{i=1}^{n} q_i^2} = \sqrt{np(1-p)}.$$

The problem is solved.

Comparing the mean number of substandard articles np with this magnitude we see that for large values of n the latter is much smaller than the former and only constitutes its small fraction. Thus, if $n = 60{,}000$, $p = 0.04$,

$$np = 2400,\ Q = \sqrt{60{,}000 \cdot 0.04 \cdot 0.96} = 48.$$

The actual number of rejects will deviate from its mean value by approximately 5% **[2%]**.

Example 2. A mechanism consists of n articles joined successively along an axis. The lengths of each can somewhat deviate from standard and they are therefore random variables supposed independent. The mean lengths of the articles and their mean square deviations are

lengths: $a_1, a_2, \ldots, a_n$; deviations: $q_1, q_2, \ldots, q_n$.

These magnitudes for the entire chain of the articles are

$a = a_1 + a_2 + \ldots + a_n$, $q = Q$ from (10.3)

so that, if $n = 9$, $a_1 = a_2 = \ldots = a_9 = 10$ *cm*, $q_1 = q_2 = \ldots = q_9 = 0.2$ *cm*, we will have $a = 90$ *cm* and $q = \sqrt{9 \cdot 0.2^2} = 0.6$ *cm*.

The length of each article deviated from its mean value by ca. 2%, but the length of the chain only deviated from its mean value by ca. 2/3%. This decrease of the relative error which occurs in the sum of random variables plays an essential role when precise mechanisms are assembled. Without such mutual compensation the assembling would have often been unsuccessful: the total length of the articles would have been either shorter or longer than necessary. Shortening the tolerated error in the lengths of the articles is inexpedient since a comparatively small increase in the precision of these lengths leads to an essential increase in the cost of the articles[36].

Example 3. A magnitude is measured n times under invariable conditions. The results of the measurements will generally differ due to *random errors* depending on the state of the instrument and observer and variations in the state of the surrounding air.

Denote the results of measurements by $\xi_1, \xi_2, \ldots, \xi_n$ assumed as usually independent and their common mean value by $\overline{\xi}$. It is natural to suppose that the mean square deviations also coincide (and equal q). The arithmetic mean of the results of measurement η is a random variable. By the addition rule

$$\overline{\eta} = \frac{1}{n}\text{m.v. } (\xi_1 + \xi_2 + \ldots + \xi_n) = \frac{1}{n}(\overline{\xi}_1 + \overline{\xi}_2 + \ldots + \overline{\xi}_n) = \overline{\xi}.$$

In essence, it was obvious from the beginning that this mean value coincides with that for each measurement. Now, by the addition rule for variances (10.3) the mean square deviation of the sum of the ξ_i is

$$Q = \sqrt{nq^2} = q\sqrt{n}$$

and the mean square deviation of η (which is equal to Q/n) is $q/\sqrt{n}$.

We have arrived at a very important conclusion: The arithmetic mean of independent and identically distributed random variables has

1) mean value: equal to that of each summand.

2) mean square deviation: $\sqrt{n}$ times smaller than that of each summand

Suppose that the mean value of the measured distance is 200 *m* and the mean square deviation of the measurements is 5 *m*. The arithmetic mean η of 100 measurements[37] will naturally have as its mean value the same distance 200 *m* but its mean square deviation will be $\sqrt{100}$ = 10 times smaller than that of a separate measurement, 0.5 *m*. We thus have grounds for expecting that the arithmetic mean of 100 measurements will be considerably nearer to the mean value 200 *m* than the result of some measurement.

The scattering of the arithmetic mean of a large number of independent magnitudes is many times less than it is for each of those magnitudes.

Chapter 11. The Law of Large Numbers

11.1. The [Bienaymé –] Chebyshev Inequality. We have repeatedly stated that the knowledge of some mean deviation of a random variable (for example, its mean square deviation) allows us to form an approximate idea about the expected largest deviations of that variable from its mean value. This remark does not yet contain any quantitative estimates, does not ensure even an approximate calculation of the probabilities of large deviations.

The following simple consideration due to Chebyshev makes all this possible. We issue from the variance of a random variable ξ (§ 10.2.2)

$$Q_\xi^2 = \sum_{i=1}^{k} (x_i - \overline{\xi})^2 p_i.$$

Let α be any positive number. Neglecting all terms of that sum in which $|x_i - \overline{\xi}| \leq \alpha$ we can only decrease it:

$$Q_\xi^2 \geq \alpha^2 \sum_{|x_i - \overline{\xi}| > \alpha} (x_i - \overline{\xi})^2 p_i.$$

The sum will decrease still more if we replace $(x_i - \overline{\xi})^2$ in each of its terms by a smaller magnitude α^2:

$$Q_\xi^2 > \sum_{|x_i - \overline{\xi}| > \alpha} p_i.$$

In the right side we have now the sum of the probabilities of those values x_i of ξ which deviate from $\overline{\xi}$ by more than α in either direction. According to the addition rule it is the probability that ξ will take one of those values. In other words, it is the probability $P(|\xi - \overline{\xi}| > \alpha)$ that the actual deviation will be larger than α. We thus have

$$P(|\xi - \overline{\xi}| > \alpha) < \frac{Q_\xi^2}{\alpha^2}. \qquad (11.1)$$

This is the **[Bienaymé –]** Chebyshev] inequality. It estimates the probability of deviations larger than any arbitrary α if only the mean square deviation Q_ξ is known. True, the estimate is often very rough[38] but sometimes it can be nevertheless applied, whereas its theoretical importance is extremely essential.

At the end of § 10.3 we considered the following example. The mean value of measurements is 200 *m*; the mean square deviation of a measurement is 5 *m*. The probability of a deviation larger than 3 *m* was very noticeable, perhaps higher than 1/2, but its exact value can certainly only be calculated when the law of distribution of the results of measurements is completely known.

We saw, however, that the mean square deviation of the arithmetic mean, η, of 100 measurements was only 0.5 *m*. The inequality (11.1) will provide

$$P(|\eta - 200| > 3) < \frac{0.5^2}{3^2} = \frac{1}{36} \approx 0.03.$$

And so, this probability is very low; actually, it is still much lower and can be practically ignored.

In Example 1 of § 10.3 we estimated the number of substandard articles (2400 with mean square deviation 48) out of 60,000. The **[Bienaymé –]** Chebyshev inequality provides the probability of the number of rejects *m* contained, say, in the interval [2300, 2500] or $|m - 2400| \leq 100$:

$$P(|m - 2400| \leq 100) = 1 - P(|m - 2400| > 100) > 1 - 48^2/100^2 \approx 0.77.$$

The actual probability is much higher.

11.2. The Law of Large Numbers. Suppose we have *n* independent variables $\xi_1, \xi_2, \ldots, \xi_n$ with the same mean value $a = 100$ *m* and the same mean square deviation $q = 5$ *m*. The mean value of their arithmetic mean η is *a*, and its mean square deviation is $q/\sqrt{n}$ (§ 10.3, Example 3). For any positive α the **[Bienaymé –]** Chebyshev inequality then leads to

$$P(|\eta - a| > \alpha) < q^2/\alpha^2 n. \qquad (11.2)$$

Then

$$P(|\eta - 200| > \alpha) < 25/\alpha^2 n.$$

We may choose a very small α, for example, $\alpha = 0.5$ *m*. Then

$$P(|\eta - 200| > 0.5) < 100/n.$$

For a very large *n* the right side is arbitrarily small; for $n = 10{,}000$ it equals 0.01 and

$$P(|\eta - 200| > 0.5) < 0.01.$$

If the probability of such unlikely events is neglected, we may state that the arithmetic mean of 10,000 measurements will almost certainly deviate from 200 *m* not more than by 50 *cm* in either direction. When desiring to shorten that deviation to 10 *cm*, we will have to choose $\alpha = 0.1$ *m*. Then

$$P(|\eta - 200| > 0.1) < \frac{25}{0.01n} = \frac{2500}{n}$$

and *n* should now be 250,000 rather than 10,000.

Generally, however small is α, the right side of inequality (11.2) can be made arbitrary small, it is only necessary to have a sufficiently large *n*. And so, we may then arbitrarily decrease the right side of the inequality (10.2) and consider the inequality of contrary sense $|\eta - a| \leq \alpha$ to be satisfied as near to certainty as desired.

If random variables $\xi_1, \xi_2, \ldots, \xi_n$ *are independent and have the same mean value a and the same mean square deviation, their arithmetic mean will be arbitrarily near to a with probability arbitrarily near to unity* (*practically certainly so near*).

This is the simplest case of the so-called *law of large numbers*, of one of the most important fundamental theorems of probability theory. It was the great Russian mathematician Chebyshev who discovered this case in the mid-19th century as a generalization of the Bernoulli theorem (§ 6.1)[39].

An isolated random variable can (as we know) often take values far apart from its mean value (can often considerably scatter) but the arithmetic mean of a large number of random variables behaves quite differently. Its scatter is not significant and with a dominant probability it only takes values very near to its mean value. This certainly occurs since the random deviations from that mean in either direction cancel each other and in most cases the summary deviation is small. And this is indeed the profound essence of that law of large numbers.

The just proved Chebyshev theorem is often utilized for judging the quality of a homogeneous material by its comparatively small sample. Thus, the quality of cotton in a boll is judged by a few of its wisps taken randomly from different parts of the boll. Similarly judged are large quantities of wheat[40]. Such judgements are highly precise. Indeed, the sample of wheat is small as compared with the whole amount of it, but it contains a large number of grains and, according to the law of large numbers, allows us to judge sufficiently precisely the mean weight of a grain and therefore the quality of the whole amount of wheat. And a boll of cotton weighing about 330 *kg* is judged by a few hundred fibres only weighing about a tenth of a gram.

11.3. The Proof of the Law of Large Numbers. Until now, we only considered the case in which all the variables $\xi_1, \xi_2, \ldots$, had the same mean value and the same mean square deviation. However, the law of large numbers is applicable under more general assumptions.

We will now study the case in which their mean values can be arbitrary (and denote them by $a_1, a_2, \ldots$), in general differing from each other. Then the mean value of the arithmetic mean η of ξ_i will be

$$A = (1/n)(a_1 + a_2 + \ldots + a_n)$$

and by the inequality (11.1) for any positive α

$$P(|\eta - A| > \alpha) < Q_\eta^2/\alpha^2. \qquad (11.3)$$

All is thus reduced to estimating Q_η^2 which is almost as simple as in the previous particular case. This magnitude is the variance of η equal

to the sum of n mutually independent variables (this is still our assumption) divided by n. By the addition rule for variances we have

$$Q_\eta^2 = \frac{1}{n^2}(q_1^2 + q_2^2 + ... + q_n^2)$$

where q_1, q_2, … are the mean square deviations of ξ_1, ξ_2, …

Now we suppose that in general these deviations also differ from each other provided however that, taking as many of them as we wish (so that n can be arbitrarily large), all of them are still smaller than some positive number b. Actually, this requirement is invariably met since we have to add magnitudes of similar, in a sense, magnitudes and the extents of their scatter do not differ too much.

And so, let $q_i < b$, $i = 1, 2, \ldots$ Then the equality above leads to

$$Q_\eta < \frac{1}{n^2} nb^2 = \frac{b^2}{n}.$$

By the inequality (11.3) we have

$$P(|\eta - A| > \alpha) < \frac{b^2}{n\alpha^2}.$$

However small is α, a sufficiently large number of the random variables will ensure that the right side of this inequality can be made as small as desired which obviously proves the law of large numbers in the present general setting.

If, therefore, a sufficiently large number n of random variables ξ_1, ξ_2, … are independent and their mean square deviations remain smaller than some positive number, the absolute expected deviations of the arithmetic mean of the variables from the arithmetic mean of their mean values can be as small as desired.

This is indeed the law of large numbers in Chebyshev's general formulation. It is important to note an important circumstance. When repeating measurements of some magnitude a under invariable conditions the observer gets not quite the same numerous results ξ_1, ξ_2, …, ξ_n and assumes that the approximate value of a is their arithmetic mean. Can we expect to obtain an arbitrarily precise value of a after carrying out a sufficiently large number of observations?

Yes. We can if only there are no systematic errors[41], if

$$\overline{\xi}_k = a, k = 1, 2, \ldots, n$$

and if the obtained values ξ_k are not indefinite; that is, if we correctly read the results on our instrument. If, however, the possible precision of reading is only δ, then, obviously, we cannot expect to obtain results more precise than $\pm\,\delta$ and the arithmetic mean of the results will be certainly corrupted by the same uncertainty[42].

This remark means that, if the instrument provides the results of observation to within some indefinite δ, the attempts to obtain the

value of *a* more precisely by applying the law of large numbers will be deceptive and the pertinent calculations become an arithmetical childish occupation.

Chapter 12. The Normal Laws

12.1. Formulation of the Problem. We have seen that some random variables essentially influence a considerable number of natural phenomena and technological processes and operations. It often occurs that until the end of a phenomenon, process or operation we can only **[?]** know the laws of distribution of these variables, i. e., the lists of their values and corresponding probabilities.

If a variable can take infinitely many different values (the range of a fired shell, the error of measurement) it is preferable to indicate the probability of some intervals of those values rather than the values themselves. For example, it is advantageous to say that that error is contained within interval [– 1, 1] or [0.1, 0.25] millimetres.

Had we wished to find out the laws of distribution of the encountered random variables[43] without taking into account general considerations or guesstimates, had we without any preliminary assumptions attempted to discover all the features of those laws by approaching each random variable purely experimentally, our problem would have been too laborious and hardly feasible. Establishing at least the most important features of a new, unknown law of distribution would have required a large number of trials. Long since scientists have therefore attempted to discover such general types of laws which could have been easily foreseen, expected, suspected to describe at least a wide class of practically encountered random variables. Long ago such types have been theoretically established and their existence experimentally confirmed.

It is obvious how advantageous is the possibility of foreseeing, by issuing from theoretical considerations and the entire previous experience, the type of the laws of distribution which necessarily describe an encountered random variable. If such guessing is confirmed, a very few trials or observations are usually sufficient for determining all the necessary features of the sought law of distribution.

Theoretical studies have shown that in a large number of cases we may with sufficient grounds expect laws of distribution of a certain type. These laws are called *normal*. Owing to the complexity involved, we briefly describe them here omitting all the proofs and exact formulations.

Among practically occurring random variables very many are *random errors* or at least are easily treated as such. Take for example the distance ξ travelled by a fired shell. We naturally assume that there exists some typical mean distance ξ_0 set as the required range. The difference $\xi - \xi_0$ is the *error* of the distance, and the study of the random variable ξ is completely and immediately reduced to studying that *random error*.

Such errors, however, change their magnitude from one shot to another. As a rule, they depend on many causes acting independently from each other: random fluctuations of the gun tube **[?]**, an unavoidable (although small) scattering of the weight and form of the shell, random changes in atmospheric conditions, random errors of aiming, – all these and still many other causes are capable of leading to error in the distance[44]. All the particular errors are mutually

independent random variables, such that *the effect of each only constitutes a very small fraction of their joint action*.

The final error $\xi - \xi_0$ which we desire to study will simply be the summary effect of all the separate mutually independent random errors. A similar situation clearly exists for most practically encountered random errors. Theoretical considerations show that the law of distribution of a random variable which is the sum of a very large number of mutually independent random variables of whichever essence, *if only* **[the action of]** *each of them is small as compared with* **[that of]** *the whole sum* ought to be near to the law of a completely determined type, the type of normal laws[45].

We are thus able to assume that a very considerable part of practically encountered random variables (in particular, all those caused by a large number of mutually independent errors) are distributed approximately according to normal laws. We ought therefore to acquaint ourselves with their main features.

12.2. Notion of Curves of Distribution. Laws of distribution can be advantageously shown on diagrams. They allow us to see at a glance, without studying any tables, the most important features of those laws. The possible values of a given random variable ξ are marked by points on a horizontal line beginning from some point of origin, positive values to the right and negative, to the left. The probabilities of each such value are marked upwards along perpendiculars erected at the points corresponding to those values. The scales in both directions are chosen in a manner that ensures a convenient and easily visible picture.

By the addition rule, the probability that ξ takes a value contained within some interval (α, β) equals the sum of the probabilities of all such possible values. If, as it often happens, the number of these values is very large, the top points of the corresponding perpendiculars seem merged into a single continuous curve, the *curve of distribution* of the studied random variable. The probability of the inequalities $\alpha < \xi < \beta$ is represented by the sum of the lengths of the perpendiculars located within the interval (α, β).

Suppose that the distance between two adjacent possible values of the random variable is always unity if, for example, those values are expressed by successive integers. This we can always actually attain by selecting an appropriate scale for our diagram. The length of each perpendicular will then be numerically equal to the *area of a rectangle* whose height is that very length and the base is the unit distance between adjacent possible values of the random variable.

It is easy to understand that the probability of the inequalities $\alpha \leq \xi < \beta$ can be represented by the sum of such rectangles situated above segment $[\alpha, \beta]$. Practically, however, if those possible values are very densely disposed, that sum will not differ from the area of a curvilinear figure bordered by the segment $[\alpha, \beta]$ from below, by that figure from above and, from the sides, by the perpendiculars erected from α and β. The probability of the studied random variable to fall in any interval is simply and conveniently given by the area above that interval and below the curve of distribution.

As a rule, when a random variable takes very many possible values the probabilities of separate values are negligible (practically zero) and

uninteresting. Thus, when measuring the distance between two settlements, it is utterly uninteresting to know that its error is exactly 473 *cm*. On the contrary, of essential interest is the probability of a deviation contained between 3 and 5 *m*[46]. The same is true in all similar cases: when a random variable takes very many values, it is important to know the probability of intervals of those values rather than of separate values.

12.3. Properties of the Curves of Normal Distributions. A normally distributed variable always takes infinitely many possible values. In spite of all the differences between normal curves they have common pronounced features:

1) All those curves have a single peak and incessantly drop on its both sides. When removing an interval of possible values of a random variable in either direction from the perpendicular of that peak the probability that that variable takes a value within the interval will continuously lower.

2) All those curves are symmetric with regard to the perpendicular passing through that peak. The areas situated above segments of equal areas and equally removed from that perpendicular are therefore obviously equal.

3) All those curves are bell-shaped. In the vicinity of the peak they are convex upward, then, at some distance from the peak they inflect and become convex downward. That distance (and the height of the peak as well) differ for different curves[47].

So how do the various normal curves differ from each other? When answering this question, we ought to recall first of all that the complete area between any curve of distribution **[not only normal]** and the chosen horizontal line is unity since it equals the probability that the given random variable takes any of its values, – equals the probability of a certain event.

The difference between curves of distribution only consists in the difference in which that summary area, the same for all of them, is distributed along that horizontal line. For normal curves the main question is, how much of that summary area is concentrated above intervals adjacent to the perpendicular of the peak and how much above more remote intervals. If almost all this summary area is concentrated in the vicinity of the peak, the random variable will with overwhelming probability (and therefore in an overwhelming number of cases) take values near it. Such variables are little scattered and their variances are small. Owing to the symmetry of the normal curve the most probable value of the random variable always coincides with its mean value.

If, on the contrary, only a small part of all this summary area is concentrated in the vicinity of the peak, the random variable will likely take values notably deviating from its most probable value. Such variables are much scattered and their variances are large.

For acquainting ourselves most rapidly with all the totality of the normal laws and learning how to apply them it is expedient to issue from their main properties.

Main Property 1. *If* ξ *is distributed according to a normal law, then*

for any constant $c > 0$ and d the variable $c\xi + d$ is also distributed according to some normal law; and, conversely, for any normal law there exists such a (*unique*) *pair of numbers $c > 0$ and d that $c\xi + d$ is distributed according to that very law*.

And so, if random variable ξ has a normal distribution, all the laws of distribution of $c\xi + d$ for any $c > 0$ and d are also normal.

Main property 2. *If two random variables are independent and distributed according to normal laws, their sum is also distributed according to some normal law*.

We can now rigorously justify some **[other]** properties especially important for applications.

1) *For any two numbers a and $q > 0$ there exists a unique normal law with mean value a and mean square deviation q*.

Indeed, let ξ be a normally distributed random variable with mean value $\overline{\xi}$ and mean square deviation Q_ξ. By Main Property 1 this statement will be proved if we show that there exists such a unique pair of numbers $c > 0$ and d that $c\xi + d$ has mean value a and mean square deviation q. Suppose that ξ takes a finite number of values. Then we may reason in the following way. Let the law of distribution of ξ be

values: $x_1, x_2, \ldots, x_n$; probabilities: $p_1, p_2, \ldots, p_n$.

The variable $c\xi + d$ (where $c > 0$ and d are yet any constants) will have the following law of distribution:

values: $cx_1 + d, cx_2 + d, \ldots, cx_n + d$; probabilities: $p_1, p_2, \ldots, p_n$.

Obviously[48],

$$\sum_k x_k p_k = \overline{\xi},\ \sum_k (x_k - \overline{\xi})^2 p_k = Q_\xi^2.$$

We ought to prove that

$$\sum_k (cx_k + d)p_k = a,\ \sum_k (cx_k + d - a)^2 p_k = q^2.$$

The first equality leads to

$$c\sum_k x_k p_k + d\sum_k p_k = a,\ c\overline{\xi} + d = a \qquad (12.1)$$

and the second provides

$$\sum_k (cx_k + d - c\overline{\xi} - d)^2 p_k = c^2 \sum_k (x_k - \overline{\xi})^2 p_k = c^2 Q_\xi^2 = q^2.$$

Therefore, since $c > 0$,

$$c = q/Q_\xi \qquad (12.2)$$

and, by (12.1)

$$d = a - c\overline{\xi} = a - \frac{q\overline{\xi}}{Q_\xi}. \qquad (12.3)$$

Formulas (12.2) and (12.3) also persist for random variables taking infinitely many values.

And so, given a and q, numbers c and d can always be uniquely determined by those formulas and $c\xi + d$ obeys the normal law with mean value a and mean square deviation q, QED.

If we consider every possible laws of distribution rather than normal laws, the knowledge of the mean value and variance (or mean square deviation) of a random variable will yet offer very little information about its law of distribution. There exists a lot of laws of distribution (and for that matter essentially differing from each other) having the same mean values and the same variances. In general, the knowledge of those magnitudes only briefly characterizes a law of distribution.

The situation is different if we restrict our attention to normal laws. On the one hand, as we saw just above, any assumption about the mean value and variance of a given random variable is compatible with its obeying a normal law. On the other hand, and this is the main point, if we have grounds for assuming beforehand that a variable obeys some normal law, that law is uniquely determined by the knowledge of these mentioned parameters, and its essence as a random variable is completely established. In particular, we can calculate the probability that its value belongs to some arbitrarily chosen interval.

2) *The ratio of the probable to the mean square deviation is the same for all normal laws.*

Suppose that we are given two arbitrary normal laws with ξ obeying the first of them. By Main Property 1 there exist such constants $c > 0$ and d that $c\xi + d$ obeys the second of these laws. Denote the mean square deviation and the probable deviation by Q_ξ and E_ξ respectively for the first variable and by q and ε for the second. By definition of the probable deviation

$$P(|(c\xi + d) - \text{m. v. } (c\xi + d)| < \varepsilon) = 1/2 \text{ or } P(c|\xi - \overline{\xi}| < \varepsilon) = 1/2 \text{ or}$$
$$P(|\xi - \overline{\xi}| < \varepsilon/c) = 1/2.$$

And again by that definition ε/c is the probable deviation of ξ: $\varepsilon/c = E_\xi$ and $\varepsilon/E_\xi = c$. Therefore (12.2) leads to

$$\varepsilon/E_\xi = q/Q_\xi \text{ and } \varepsilon/q = E_\xi/Q_\xi.$$

The chosen normal laws were arbitrary which means that the formulated proposition is proved. The ratio ε/q is an absolute constant **[not depending on the choice of the normal law]**; denote it by λ. It is known that $\lambda = \sqrt{2/\pi} \approx 0.674$ which means that for any normal law $\varepsilon = q\sqrt{2/\pi}$.

Because of this extremely simple connection between ε and q for normally distributed variables the choice of one or another for characterizing scatter is actually indifferent. It was stated above (even without restricting our study to normal laws) that unlike other characteristics the mean square deviation has many simple properties which in most cases compels both theoreticians and practitioners to choose those very deviations as a measure of scatter.

We have also remarked that artillery men nevertheless almost always apply probable deviations but we see now why this tradition is harmless. Random variables, with which the theory and practice of artillery firing are dealing, are almost always normally distributed. For such variables, because of the proportionality involved, the choice of any of those two characteristics is practically indifferent.

3) *Suppose that* ξ *and* η *are independent normally distributed random variables and* ς = ξ + η. *Then*

$$E_\varsigma = \sqrt{E_\xi^2 + E_\eta^2}$$

where E_ς, E_ξ *and* E_η *are the probable deviations of the corresponding variables.*

We (§ 10.3) know that a similar formula takes place for mean square deviations whichever are the laws of distribution of ξ and η. For normally distributed ξ and η the variable ς is also normally distributed (Main Property 2) and by property **2)**,

$$E_\xi = \lambda Q_\xi,\ E\eta = \lambda Q_\eta,\ E\varsigma = \lambda Q_\varsigma,$$

$$E\varsigma = \lambda\sqrt{Q_\xi^2 + Q_\eta^2} = \sqrt{(\lambda Q_\xi)^2 + (\lambda Q_\eta)^2} = \sqrt{E_\xi^2 + E_\eta^2}.$$

For normal laws, one of the most important properties of mean square deviations thus directly extends to probable deviations.

12.4. Problems and Examples. We will call a normal distribution *standard normal* if its mean value is 0 and its variance, 1. For the sake of brevity a random variable ξ obeying this law is written as

$$P(|\xi| < a) = \Phi(a),\ a > 0.$$

Φ(a) is thus the probability that the absolute value of ξ is less than a. Very precise tables of Φ(a), irreplaceable for those who calculate probabilities, have been compiled and are appended to each book devoted to probability, – to this book as well [not reproduced here]. All calculations with any normal variable can be easily and very precisely carried out by means of such tables. We show now how this is done.

Problem 1. Random variable ξ is normally distributed with mean value $\overline{\xi}$ and mean square deviation Q_ξ. Required is the probability that the absolute deviation $|\xi - \overline{\xi}|$ is less than a **[$a > 0$]**.

Let ς be a random variable distributed according to the standard normal distribution. By Main Property 1 there exist such numbers

$c > 0$ and d that $c\varsigma + d$ has mean value $\overline{\xi}$ and mean square deviation Q_ξ. In other words, that it has the same law of distribution as ξ. Therefore

$$P(|\xi - \overline{\xi}| < a) = P(|(c\varsigma + d) - (c\overline{\varsigma} + d)| < a) = P(c|\varsigma - \overline{\varsigma}| < a).$$

However, by formula (12.2) $c = Q_\xi/Q_\varsigma = Q_\xi$ since $Q_\varsigma = 1$ because of the standard normal distribution of ς. Therefore

$$P(|\xi - \overline{\xi}| < a) = P(Q_\xi|\varsigma - \overline{\varsigma}| < a) = P(|\varsigma| < \frac{a}{Q_\xi}) = \Phi(\frac{a}{Q_\xi}). \qquad (12.4)$$

The problem is solved since $\Phi(a/Q_\xi)$ can be directly found in a table. For variables obeying any normal distribution our table thus allows us to calculate easily by formula (12.4) the probability of any boundary of the deviations of a variable obeying any normal law from its mean value.

Example 1. A certain article is manufactured on a lathe. Its length ξ is a normally distributed random variable with mean value 20 *cm* and variance 0.2 *cm*. Required is the probability that that length will be contained within 19.7 and 20.3 *cm*, – that its deviation in either direction will be less than 0.3 *cm*.

By formula (12.4) and our table

$$P(|\xi - 20| < 0.3) = \Phi(0.3/0.2) = \Phi(1.5) = 0.866.$$

The lengths of about 87% of the articles will be contained between 19.7 and 20.3 *cm*. The length of the other articles will deviate more than by 0.3 *cm* from the mean value.

Example 2. Keeping to the conditions of Example 1, find the precision of the length of an article that can be guaranteed with probability 95%.

We obviously have to find such a positive number a for which

$$P(|\xi - 20| < a) > 0.95.$$

We saw just above that the value $a = 0.3$ is too small since the left side of the new inequality will then be less than 0.87. According to formula (12.4)

$$P(|\xi - 20| < a) = \Phi(a/0.2) = \Phi(5a).$$

Therefore, we ought to determine first of all **[?]** such a value of $5a$ for which $\Phi(5a) > 0.95$. Our table provides $5a > 1.97$ and $a > 0.394 \approx 0.4$ (*cm*).

Example 3. In practice, it is sometimes assumed that a normally distributed random variable ξ does not deviate **[from its empirical (!) mean]** more than by three mean square deviations. What grounds do we have for that assumption?

Formula (12.4) and our table show that

$$P(|\xi - \overline{\xi}| < 3Q_\xi) = \Phi(3) > 0.997,\ P(|\xi - \overline{\xi}| > 3Q_\xi) < 0.003.$$

This actually means that larger deviations will in the mean occur rarer than 3 times in a thousand. May we neglect this possibility or should it be necessarily taken into account? The answer certainly depends on the essence of the problem at hand and cannot be provided once and for all. Note also that the relation

$$P(|\xi - \overline{\xi}| < 3Q_\xi) = \Phi(3)$$

is a particular case of the formula

$$P(|\xi - \overline{\xi}| < aQ_\xi) = \Phi(a) \qquad (12.5)$$

which follows from (12.4) and takes place for any normally distributed random variable ξ.

Example 4. The mean weight of a certain article is 8.4 *kg*. It is found that absolute deviations larger than 50 *g* occur in the mean 3 times out of a hundred. Assume that the weight is normally distributed and determine its probable deviation.

Given,

$$P(|\xi - 8.4| > 0.05) = 0.03$$

where ξ is the weight of a randomly chosen article. Therefore

$$0.97 = P(|\xi - 8.4| > 0.05) = \Phi(0.05/Q_\xi).$$

The table shows that $\Phi(a) = 0.97$ at $a \approx 2.12$. Therefore

$$0.05/Q_\xi \approx 2.12,\ Q_\xi \approx 0.05/2.12.$$

The probable deviation is, see § 12.3,

$$E_\xi = 0.674Q_\xi \approx 0.0155\ kg = 15.5\ g.$$

Example 5. Deviations of a shell fired from an artillery gun result from three mutually independent causes: errors of determining the location of the target; of aiming; from causes changing from one shot to another (the weight of the shell, atmospheric conditions etc)[49]. Assuming that all these errors are normally distributed with mean values 24, 8 and 12 *m* respectively, determine the probability that the summary deviation from the target will not exceed 40 *m*.

By property **3)** the probable deviation of the summary error ξ is

$$\sqrt{24^2 + 8^2 + 12^2} = 28\ (m)$$

so that the mean square deviation of that error is 28/0.647 **[0.674]** ≈ 41.5,

$$P(|\xi| < 40) = \Phi(40/41.5) \approx \Phi(0.964) = 0.665.$$

Deviations not larger than 40 *m* thus occur in approximately 2/3 of cases.

Problem 2. Random variable ξ is normally distributed with mean value $\overline{\xi}$ and mean square deviation Q_ξ. Required is the probability that the absolute deviation $|\xi - \overline{\xi}|$ is contained within interval $[a, b]$.

By the addition rule

$$P(|\xi - \overline{\xi}| < b) = P(|\xi - \overline{\xi}| < a) + P(a < |\xi - \overline{\xi}| < b),$$
$$P(a < |\xi - \overline{\xi}| < b) = P(|\xi - \overline{\xi}| < b) - P(|\xi - \overline{\xi}| < a) =$$
$$\Phi(b/Q_\xi) - \Phi(a/Q_\xi). \qquad (12.6)$$

The problem is thus solved.

In an overwhelming majority of practical requirements the table of $\Phi(a)$ which we have used all the time is however an excessively awkward tool. Usually it is only needed to calculate the probability of the deviation $\xi - \overline{\xi}$ contained within more or less long intervals. It is therefore desirable to have, along with our *complete* table, an abbreviated table. Such tables are easy to compile from a complete table by means of formula (12.6). Here is an example; the abbreviated table is much less precise than the table appended here, but it still is quite sufficient for many cases.

Separate the entire range of magnitude $|\xi - \overline{\xi}|$ into five parts

from 0 to $0.32Q_\xi$; from $0.32Q_\xi$ to $0.69Q_\xi$; from $0.69Q_\xi$ to $1.15Q_\xi$; from $1.15Q_\xi$ to $2.58Q_\xi$; and beyond that.

By formula (12.4) we have

$$P(|\xi - \overline{\xi}| < 0.32Q_\xi) = \Phi(0.32) \approx 0.25.$$

Similar calculations provide the other probabilities. They are approximately equal to 0.25; 0.25; 0.24; and 0.01. The entire infinite axis can be separated into 10 intervals, five of them positive **[in the positive semi-axis]** and five negative. We will then immediately imagine the main features of the distribution of the deviations of the random variable with both arbitrary parameters.

Finally, we consider the calculation of probabilities of a normally distributed random variable to be contained within an arbitrary interval.

Problem 3. Random variable ξ is normally distributed with mean value $\overline{\xi}$ and mean square deviation Q_ξ. It is required to calculate by means of a table the probability of inequalities $a < \xi < b$, $a < b$. Both these numbers are arbitrary.

We have to study three cases depending on the arrangement of *a* and *b* with respect to $\overline{\xi}$. Note that for any normally distributed random variable and any number *c* the probability of the equality $\xi = c$ is zero.

First case: $\overline{\xi} \le a \le b$. By the addition rule

$$P(\overline{\xi} < \xi < b) = P(\overline{\xi} < \xi < a) + P(a < \xi < b),$$

$$P(a < \xi < b) = P(\overline{\xi} < \xi < b) - P(\overline{\xi} < \xi < a) =$$
$$P(0 < \xi - \overline{\xi} < b - \overline{\xi}) - P(0 < \xi - \overline{\xi} < a - \overline{\xi}).$$

However, because of the symmetry of the normal laws, for any $\alpha > 0$

$$P(0 < \xi - \overline{\xi} < \alpha) = P(-\alpha < \xi - \overline{\xi} < 0) = 1/2P(-\alpha < \xi - \overline{\xi} < \alpha) =$$
$$1/2P(|\xi - \overline{\xi}| < \alpha) = 1/2\Phi(\alpha/Q_\xi). \qquad (12.7)$$

Therefore $P(a < \xi < b) = \frac{1}{2}[\Phi(\frac{b-\overline{\xi}}{Q_\xi}) - \Phi(\frac{a-\overline{\xi}}{Q_\xi})]$.

Second case: $a \le \overline{\xi} \le b$. By the addition rule

$$P(a < \xi < b) = P(a < \xi < \overline{\xi}) + P(\overline{\xi} < \xi < b) =$$
$$P(a - \overline{\xi} < \xi - \overline{\xi} < 0) + P(0 < \xi - \overline{\xi} < b - \overline{\xi}) =$$
$$\frac{1}{2}[\Phi(\frac{\overline{\xi}-a}{Q_\xi}) + \Phi(\frac{b-\overline{\xi}}{Q_\xi})],$$

see formula (12.7).

Third case: $a \le b \le \overline{\xi}$. By the addition rule

$$P(a < \xi < \overline{\xi}) = P(a < \xi < b) + P(b < \xi < \overline{\xi}),$$

$$P(a < \xi < b) = P(a < \xi < \overline{\xi}) - P(b < \xi < \overline{\xi}) =$$
$$P(a - \overline{\xi} < \xi - \overline{\xi} < 0) - P(b - \overline{\xi} < \xi - \overline{\xi} < 0) =$$
$$\frac{1}{2}[\Phi(\frac{\overline{\xi}-a}{Q_\xi}) - \Phi(\frac{\overline{\xi}-b}{Q_\xi})].$$

The problem is completely solved. We see that for a random variable distributed according to any normal law our table allows us to calculate the probability of this variable to be contained within any interval and thus to characterize exhaustively its law of distribution. The following example shows how to achieve this.

Example 6. Shells are fired from point O along straight line OX. The mean distance travelled by the shells *H* is 1200 *m*. Suppose that that distance is normally distributed with mean square deviation 40 *m*. Determine the per cent of overshots contained within 60 – 80 *m*.

We are determining the probability of $1260 < H < 1280$. Applying the final formula of Example 3 we find

$$P(1260 < H < 1280) = \frac{1}{2}[\Phi(\frac{1280-1200}{40}) - \frac{1}{2}[\Phi(\frac{1260-1200}{40})] =$$

$$\frac{1}{2}[\Phi(2) - \Phi(1.5)].$$

The table provides $\Phi(2) \approx 0.955$, $\Phi(1.5) \approx 0.866$,

$P(1260 < H < 1280) \approx 0.044$.

A little more than 4% **[a little less than 4.5%]** of the shells will overshoot the target by 60 – 80 *m*.

Part 3

Stochastic Processes

Chapter 13. Introduction to the Theory of Stochastic Processes

13.1. General Idea of Stochastic Processes. When studying natural phenomena and processes occurring in technology, economics and transportation we often have to describe them by random variables changing in time. A few examples.

Diffusion is known to consist in molecules of a substance penetrating into another substance and intermingling with its molecules. Let us trace the motion of a molecule. Suppose that at initial moment $t_0 = 0$ it was in position (x_0, y_0, z_0) and the components of its velocity were (v_{0x}, v_{0y}, v_{0z}). It collides with other molecules at random moments and changes its position as well as velocity and direction of motion. It is impossible to foresee exactly this change since we do not know either the moments of the collisions or their number during any interval of time or the velocities **[or directions]** of those other molecules.

The position of a molecule at moment t is determined by three components $x(t)$, $y(t)$, and $z(t)$ which are thus random functions of time. The components of velocity $v_x(t)$, $v_y(t)$, and $v_z(t)$ are random variables changing in time as well[50].

Consider now a complicated device consisting of a large numbers of elements (capacitors, resistances, diodes, mechanical parts etc). Owing to some causes each element can loose its working properties and quit functioning. We will call such a state a *failure*. Observations of various technical devices over long periods of time are showing that the period of work from beginning to failure cannot be precisely indicated beforehand since it is a random variable.

Suppose now that as soon as some element fails it is replaced by a new element and that the work of the studied device continues. How many elements should be replaced during time interval $[0, t]$? Denote this number by $n(t)$ which depends on t and is random. This is a new example of a random variable changing in time. Its special feature is that it cannot decrease and randomly changes by integers (by the number of the elements which have to be changed). Such random functions are considerably interesting in the *theory of reliability* **[cf. § 13.5]**, an important engineering science which widely applies the methods of probability theory.

Modern industry needs electricity. How much energy will be consumed by a factory or shop during a given interval of time? How large can the consumed power be at each given moment? How to calculate the parameters of electrical cables which should not be too low of capacity and should not burn out during a period of normal work either? And the sections of these cables should not be too large, otherwise an excessive expenditure of metal becomes necessary and considerable capital is withdrawn from circulation.

Answers to these questions naturally require a thorough study of the consumption of electricity by separate lathes, mechanisms, various devices and contrivances as well as by all feeders. Such investigations had been carried out at many enterprises of different branches of industry. We provide a picture typical for the metal-working branch, but the final conclusions will be the same for other types of enterprises as well.

The periods of the work of a turning lathe alternate with periods of its idleness and, accordingly, the consumed power essentially differs. From almost zero during the dead time it sharply rises but does not remain constant. It rather undergoes considerable changes since the local heterogeneity of the treated material changes the speed of the work and the exerted effort.

In addition, the periods of work and idleness change very irregularly. On closer and more thorough examination their change proves to be random and once more we have to deal with a random function of time. The sharp fluctuations of the power consumed by a lathe are smoothed when considering a group of 10 or 20 of them.

The summary consumption of power remains random, but becomes smoother. This is essentially explained by the regularities with which we became acquainted when studying the law of large numbers. The levelling is connected with the scatter of the peaks of consumption: for a certain lathe the peak often occurs during periods of less or even minimal consumption by the other lathes.

At present, the study of the electrical load of industrial plants and towns is being ever more based on the indicated features. And the ideas, methods and mathematical machinery of probability and the theory of stationary processes (of the theory of random functions of an independent variable) are indeed widely applied for solving them.

13.2. Notion of Stochastic Processes and Their Various Types. We have come now to the definition of a stochastic process. Imagine that some random variable $\xi(t)$ depends on a continuously changing parameter t usually called time. Actually, it can mean something else as well but in an overwhelming number of cases it is indeed time.

For defining a stochastic process we ought to describe the possible values which it takes at each moment, their expected changes, the probabilities of those possible changes in time and the degree of dependence of the development of the process on its previous history. Without finding out all that we cannot at all state that we know a stochastic process. According to the general method of mathematically describing a stochastic process the functions

$$F(t_1, t_2, \ldots, t_n, x_1, x_2, \ldots, x_n) = P[\xi(t_1) < x_1, \xi(t_2) < x_2, \ldots, \xi(t_n) < x_n]$$

are *considered given for any integer positive number n and any moments* $t_1, t_2, \ldots, t_n$.

This method of describing a stochastic method is universal; in principle, it allows us to ascertain all the features of the behaviour of the process in time. However, it is very unwieldy so that for obtaining more profound results we have to isolate particular types of stochastic processes and look for pertinent analytical tools more adapted to calculations and to constructions of mathematical models of the studied phenomena.

At present, several classes of stochastic processes are isolated in connection with various real processes and their study is sufficiently advanced. The pertinent information is, however, beyond the reach of elementary mathematical knowledge. Markov processes called after the outstanding Russian mathematician Markov of the end of the 19th

and the beginning of the 20th century gained special importance. He began considering, and was the first to study systematically the properties of the so-called *chain* dependences which became the prototype for constructing the notions and theory of the Markov stochastic processes.

Suppose that process $\xi(t)$ has the following property. For any moments t_0 and t, $t_0 < t$, the probability of **[its]** passing from state x_0 at moment t_0 to state x (or to one of the states belonging to some set A) at moment t only depends on t_0, x_0, t and x (or A). Additional knowledge of the states of the process during previous periods does not change that probability. All the development of such processes as though concentrates in the state x_0 achieved at moment t_0 and only influences its further history **[is only influenced]** through that x_0. Such processes are indeed called after Markov.

At a glance, it may seem that such a serious schematization of phenomena has little in common with real requirements since the after-effect of the previous development usually continues for a rather long time. However, mathematics and its applications in biology, technology, physics and other branches of knowledge had accumulated experience that shows that many phenomena such as diffusion or the management of the automatic control of manufacturing perfectly conform to the pattern of Markov processes.

Moreover, it occurred that by changing the notion of *state* any stochastic process can be converted into a Markov process. This is a very serious argument favouring a wide development of their theory. Markov processes are therefore extensively applied in studies of many practical problems since they allow the application of a well developed and comparatively simple analytical means of calculation.

Consider in addition that any application of mathematical means for studying some natural phenomena or technological, economic or mental processes requires their preliminary schematization, an isolation of some peculiarities which sufficiently describe their course. True, it is now usual to discuss simulation rather than schematization. The model of phenomena which we created possesses many peculiarities. First, it is simpler than the studied phenomenon itself. Second, its initial propositions and connections are clearly formulated, a feature lacking in real processes, and especially so in economic and biological phenomena. After studying a comparatively simple model of a phenomenon and comparing the formulated conclusions with observations of the phenomenon itself, we can judge the quality of our model and specify it if necessary.

When constructing a mathematical model, it is tacitly assumed that mathematical analysis is only applicable to studying the process of the changes of some system if each of its possible states and its evolution is exhaustively described by some chosen mathematical tool. We should apparently consider the Newtonian mechanics as one of the most remarkable mathematical models of the surrounding phenomena of a certain kind.

A simple pattern of the course of a process and the connected mathematical arsenal of the differential and integral calculus have by now been perfectly describing numerous processes for a quarter of a

millennium. The advances of mechanical engineering and the flights of the first spaceships not only in the Earth's vicinity but to other planets as well are essentially based on a wide application of the classical Newtonian mechanics. It assumes that the motion of a system of mass points is completely described by the position and velocity of each of them. In other words, by indicating these data for moment *t* allows us to calculate the unique state of our system for any other moment. For achieving this aim mechanics offers equations of motion.

Note that the state of a system of points only understood as their positions at moment *t* is insufficient for uniquely determining subsequent states of the system. For the Newtonian mechanics, the **[mentioned]** notion of state ought to be extended by adding the values of the velocities at a given moment.

All that which is situated beyond classical mechanics, that is, all modern physics, has to deal with a considerably more complicated situation in which the knowledge of the state of a system at a given moment cannot anymore uniquely determine its future states. For Markov processes uniquely determined are only the probability of passing into some state during a certain period of time. We may consider Markov processes as a wide extension of processes studied by classical mechanics[51].

13.3. Simplest Flows of Events. In many practically important situations or those interesting from cognition we have to ascertain the regularities in the occurrence of certain events (of ships arriving at a seaport, failures of complicated devices, changes of burned out bulbs, moments of the decay of the atoms of a radioactive substance etc). Calculations pertaining to the work of consumer services (hairdressers, shops, public transportation, number of beds in hospitals, capacities of locks, crossings, bridges etc)[52] are closely linked with studying such flows. In the 1930s the moments of arrival of airplanes at large airports, of cargo boats at seaports, the calls to first-aid stations and telephone exchanges etc had been thoroughly studied. It occurred that in all those cases the occurrences of the mentioned events were sufficiently well described by the following regularity.

Suppose that $P_k(t)$ is the probability of the occurrence of k events of a flow during time interval t. Then, for $k = 0, 1, 2, \ldots$ the equalities

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad (13.1)$$

are satisfied with a high precision. Here, λ is a positive constant describing the *intensity* of the occurrence of the events of the flow. In particular, the probability that no event arrives during time t is

$$P_0(t) = e^{-\lambda t}. \qquad (13.2)$$

Molecular physics studies the probability that during a given period of time t a given molecule will not collide with any other molecule. Books devoted to such problems indicate that that probability indeed equals $e^{-\lambda t}$. If the flow of events is here understood as the moments of

collisions of the given molecule with other molecules we will indeed determine the probability that no event will occur during time *t*.

It is natural to suppose that there exists a general cause leading to the occurrence of the same regularity of those so differing phenomena. And it was indeed discovered that under very wide conditions there exist various and profound causes leading to the just described regularity. Already at the beginning of the 20th century Einstein and Smoluchowski who studied the Brownian motion discovered the first group of such conditions. Suppose that a flow of events has the following three properties:

1. Stationarity: *For any finite number of non-intersecting intervals of time the probability of the occurrence of k_1, k_2, …, k_n events only depends on those numbers and on the duration of the time intervals. In particular, the probability of the occurrence of k demands in interval* (*T*, *t* + *T*) *does not depend on T and is only a function of k and t.*

2. Lack of after-effect: *The probability of the arrival of k events of a flow during time interval* (*T*, *T* + *t*) *does not depend on the number of the previously arrived events or on how did they arrive.* This requirement means that the studied flow is a Markov process.

3. Ordinariness: *The occurrence of two or more events during a very short period of time is practically impossible.*

A flow of events satisfying these three conditions is *a simplest flow*. It can be proved that equation (13.1) completely characterizes a simplest flow which can also be defined otherwise: it is a flow of randomly distanced moments of time with formula (13.2) indicating the probability that the distance between adjacent moments is longer than *t*. This definition is also frequently used when solving many applied and theoretical problems.

A direct check of the fulfilment of the three mentioned conditions (stationarity, lack of after-effect and ordinariness) is often difficult and it is therefore very important to derive other conditions for deciding on other grounds whether a flow is simplest or near to being it. A number of researchers have found such a condition, and here it is.

Suppose that the studied flow is a sum of a very large number of stationary flows each only little influencing the sum. Add a restriction of an arithmetical nature which ensures the ordinariness of the summary flow, and it becomes *near-simplest*. This theorem, which Khinchin, one of the creators of modern theory of probability, proved in a general form, is fundamentally important for applications. Indeed, it very often ensures formulation of serious conclusions by issuing from the general structure of a flow.

Thus, a flow of calls arriving at a telephone exchange can be considered as a sum of many independent flows each insignificantly influencing that sum. It follows that that summary flow ought to be near-simplest. Just the same, a flow of cargo boats arriving at a given seaport consists of a large number of flows departing from various other seaports and should therefore be near-simplest, and so it really is. Other examples are also possible[53].

13.4. A Problem in the Queuing Theory. The following problem is typical for many practically important cases. We will first describe it in its applied aspect, as it frequently appears to designers of plants,

department stores, storehouses, and telephone exchanges.

There are various businesses and establishments for satisfying some requirements of the population: hairdressers', telephone exchanges, hospitals, dental out-patients' clinics. Demands for service arrive at random moments and the duration of services is also random. How to meet these demands if there are n servers/servicing facilities?

It is easy to see that the described picture sufficiently reflects the real situation. We are unable to indicate just when will the customers arrive at a hairdressers' or dental clinic and we know well enough that it is often necessary to wait for service but that sometimes we are serviced immediately. Just the same, the time required for completing an apparently the same operation seems to be constant, but actually considerably differs from one case to another. A treatment of a tooth can only consist in its cleaning or, alternatively, in filling it.

Both customers and managers are first of all naturally interested in such characteristics of service as the length of queues, average waiting time, traffic intensity provided that the average rates of the arrival of demands and servicing are known. We assume that

1) The flow of demands for service is simplest.

2) The duration of servicing is random and the probability of its being not less than t equals e^{-vt} with a constant positive v.

3) Each demand is served by one server/servicing facility. Each server/servicing facility services one demand at a time.

4) If a queue has formed, as soon as the server serves his customer, he begins to serve the next one.

Denote by $P_k(t)$ the probability that at moment t there are k demands. Under the stipulated conditions these probabilities can be defined for any $k = 0, 1, 2, \ldots$ However, the precise formulas are awkward and other, preferable formulas are derived from them for an established pattern of work. They are incomparably simpler:

$$p_k = \frac{\rho^k}{k!} p_0,\ 0 \le k \le n;\ p_k = \frac{\rho^k}{n! n^{k-n}} p_0,\ k \ge n. \qquad (13.3, 13.4)$$

$$p_0 = 1 \div [\sum_{k=0}^{n} \frac{\rho^k}{k!} + \frac{\rho^{n+1}}{n!(n-\rho)}],\ \rho < n;\ p_0 = 0,\ \rho \ge n. \qquad (13.5)$$

In these formulas, $\rho = \lambda/v$. By formulas (13.3) and (13.4) it occurs that at $k \ge 1$ $p_k = 0$ as well.

This means that if $\rho \ge n$ and the process of serving is established, any finite number of demands can only exist with zero probability; infinite many demands and an infinitely long queue will exist with probability 1. If $\rho \ge n$, the queue will unboundedly grow with time.

Our conclusion is very important. Since the number of servers/servicing facilities (runways in airports, berths in seaports, beds in hospitals, cash desks in shops etc) is often calculated under a false assumption of an *ideal* capacity of a system equal to the product of the number of servers/servicing facilities by the duration of their work in a given period divided by the average duration of servicing one demand [ideal traffic intensity]. Owing to the irregular arrival of

demands such calculations lead to queues and therefore to waste of time and loss of money and potential customers.

The methods of the theory of queuing certainly ensure the possibility of ascertaining the damage inflicted by overloading a system as well as the losses incurred by having excessive servers/service facilities. Many examples can be provided for showing that that theory had been necessary when devising telephone exchanges, establishing teams of repairmen in factories, planning the capacity of large airports or tunnels for highways with heavy traffic. Nowadays, the theory of queues is becoming ever more important for designing computers, search machines, in nuclear physics, biology etc.

13.5. On a Problem in the Theory of Reliability. During the last quarter of the 20^{th} century serious worldwide attention has been paid to a new scientific discipline christened *theory of reliability*. It aims at developing general rules for designing, manufacturing, accepting, transporting, storing, and applying industrial articles for ensuring maximal efficiency of their usage.

In addition, the theory of reliability naturally works out methods for calculating the reliability of complicated articles and technical systems by issuing from the characteristics of the reliability of their components. The importance of those aims is unquestionable since our entire life is directly and obliquely connected with the application of various technical devices and systems. We go to and from work by buses and trams, in our apartments we switch the light and turn taps on and off. Hospitals apply various pieces of equipment for aiding vital functions of patients. For example, after an operation on kidneys and during the period of their restoration artificial kidneys are functioning instead. Millions of passengers are yearly travelling across the world by air. And in each case we are extremely interested in an absolutely proper work of the applied technical means. Violation of this requirement can lead to fatal consequences: an airplane can crash, an artificial kidney can fail etc.

Such problems seem to have nothing in common with the theory of probability and ought to be solved by designers and the engineering staff of factories. Actually, however, this opinion is wrong. A large part of the problem connected with the study of quantitative calculations, elaboration of expedient plans of testing the quality of manufactured articles and formulation of the pertinent conclusions, determination of best schedules for preventive inspections and repairs, is incumbent on mathematicians. And it occurs that all the necessary main characteristics of the articles are of a stochastic nature. Thus, for mass articles manufactured by the same factory, from the same raw materials and under the same conditions the duration of work until failure is considerably scattered. We may quite definitely imagine this fact when recalling how sharply the working lives of electric bulbs are fluctuating. Sometimes they work faultlessly for a few years, but sometimes they have to be replaced after only several days.

Observations over long periods and numerous special experiments convincingly showed that we are unable to determine precisely the working life of an article and can only estimate the probability that it will not be shorter than a given number t. The theory of probability

thus confidently enters all the problems of the theory of reliability and provides the main methods for solving them.

Let us now consider a simple problem and only outline the necessary calculations. We do not therefore complicate our account but at the same time describe our problem clearly enough. It is well known that by no means there exist any absolutely reliable elements or articles. Each element, however perfect are its properties, loses them with time. For enhancing the reliability of articles we ought to follow various paths: weaken the conditions of their work, look for better materials, new structures or layouts of connections **[!]**. One of the most usual methods for achieving this aim is the introduction of redundancies. In essence, this means that redundant elements, their sets or even whole units are included in the system and begin working just as the main elements (sets, units) fail.

For ensuring uninterrupted transportation redundant diesel and electric locomotives are kept at railway junctions. All large power stations have additional current generators, especially important power lines have auxiliary lines in parallel only partly functioning during normal conditions, and cars have spare wheels.

Suppose there are n devices which ought to function simultaneously for time t. They fail independently from each other, the system **[!]** fails if at least one device fails, and the common probability that one of them will not fail during that time is p. By the Bernoulli formula the probability of an uninterrupted work of the system is p^n.

How will this probability change if the system has m redundant working devices and fails if less than n out of $(n + m)$ of them are performing? By the addition rule the probability sought is

$$\sum_{i=1}^{m} C_{m+n}^{n+i} p^{n+i} (1-p)^{m-i}.$$

Here is a simple example. Let $n = 4$, $m = 1$ and $p = 0.9$. It is not difficult to find out that the probability of an uninterrupted work of the system was previously 0.6561 but that with only a single redundant device it becomes 0.9185, 1.5 times higher and 0.9841 with two redundant devices. This is why a single redundant current generator almost completely excludes failures of power stations. The reliability of systems increases many times over by introducing *redundancy with restoration*. Each failed component is then immediately repaired and returned in reserve.

We have only considered a simplified problem of the theory of reservation by redundant elements. Much more complicated mathematics and primarily the theory of stochastic processes are necessary for studying the same problem under real conditions. Nowadays many important problems of the theory of reliability are already solved, but a large number of them are still far from being satisfactorily and fully dealt with. Systematic work will allow their solution under somewhat weakened conditions and open the way for studying them under more real assumptions.

Conclusions

During the latest decades the theory of probability became one of the most rapidly developing mathematical sciences. New theoretical results reveal other possibilities for applying its methods in natural sciences and practice. At the same time, subtler and more detailed studies of natural phenomena, technological, economic and other processes prompt the theory of probability to search for new methods and discover new regularities generated by randomness. This theory is one of those mathematical sciences which do not cut themselves off from life or the requirements of other sciences, but are rather keeping abreast of the general development of natural sciences and technology.

The reader should not however wrongly think that the theory of probability has now only become an auxiliary means for solving applied problems. Not at all! During the latest decades it became a harmonious mathematical science with its own problems and methods of research. And it also occurred that the most important and natural problems of the theory of probability considered as a mathematical science are helping to achieve urgent aims in applied fields.

The theory of probability originated in the mid-17th century in connection with the works of Fermat (1601 – 1665), Pascal (1623 – 1662) and Huygens (1625 **[1629]** – 1695). Embryos of the notions of probability of a random event and expectation of a random variable have appeared in their work. Their starting point was the study of problems connected with games of chance, but they clearly saw the importance of the new concepts for studying nature. For example, Huygens stated[54]:

The reader will soon understand that I have thrown out the elements of a new theory, both deep and interesting.

Among scholars who had essentially influenced the development of the theory of probability it is necessary to indicate Jakob Bernoulli (1654 – 1705) already mentioned above, De Moivre (1667 – 1754), Bayes (**[ca. 1701 –]** 1763), Laplace (1749 – 1827), Gauss (1777 – 1855) and Poisson (1781 – 1840).

A powerful development of the theory of probability had been closely linked with the traditions and advances of Russian science. In the 19th century, in Europe, this theory came to a dead end whereas the Russian mathematician P. L. Chebyshev (1821 – 1894), a man of genius, discovered a new direction of its further development, a thorough study of sequences of independent random variables[55].

He himself and his students, A. L. [A. A.] Markov (1856 – 1922) and A. M. Liapunov (1857 – 1918), by following him, arrived at fundamental results (the law of large numbers, the Liapunov theorem). Readers are already acquainted with the law of large numbers and our next aim is to provide a notion of another most important proposition of the theory of probability which became known as the Liapunov theorem (or the central limit theorem).

It is important since a considerable number of phenomena whose outcomes depend on chance largely obey the following pattern: the studied phenomenon is influenced by great many independently acting random factors each of which only insignificantly affects its general

course. The action of each such factor is expressed by random variables $\xi_1, \xi_2, \ldots, \xi_n$, and their summary influence[56], by their sum,

$$S_n = \xi_1 + \xi_2 + \ldots + \xi_n.$$

It is practically impossible to take account of each (in other words, to indicate their laws of distribution) or even to enumerate them.

Clearly, therefore, the development of methods allowing the study of their summary action independently from the essence of each separate summand is of utmost importance. Usual methods of research are here helpless and ought to be replaced by those for which the large number of acting factors will be not an obstacle, but, on the contrary, a relief. Such methods should compensate the insufficient knowledge of each isolated factor by their large number.

The central limit theorem largely established by Chebyshev, Markov and Liapunov, states that, if the acting causes $\xi_1, \xi_2, \ldots, \xi_n$ are mutually independent, their number very large and the action of each as compared with their summary influence unimportant, the law of distribution of their sum S can only slightly differ from a normal law.

Here are pertinent examples. When firing shells, the unavoidable deviations of the hit-points from the target are represented by the well known phenomenon of scattering. It is the result of the influence of a great number of independently acting causes (irregular milling of some parts of a shell, irregular density of its material, insignificant variations in the standard amount of the explosive, unnoticeable errors in aiming the artillery gun, insignificant variations in the state of the atmosphere and many others) each of which only insignificantly influences the shell's (the shells') path(s). The theory of firing takes this fact into account and reflects it in manuals.

When measuring some physical magnitude, a great many factors unavoidably influence the obtained results. Taken by itself, each such factor cannot be accounted for, but they lead to errors of measurements. Among them are the changes in the state of the instrument whose indications can somewhat vary under the influence of various atmospheric, thermal, mechanical and other causes. There also are the errors of the observer caused by the peculiarities of his eyesight or hearing which also change with his mental or physical condition. The actual error of observation is thus the result of great many insignificant, mutually independent, so to say elementary errors depending on chance. By the Liapunov theorem we may again expect that the errors of observation obey a normal law[57].

Any number of such examples can be provided: the positions and velocities of gas molecules determined by a large number of collisions with other molecules; the amount of a diffused substance; deviations of the sizes of machine parts from the standard in mass manufacturing; the distribution of the heights of animals **[of the same species]** or of the sizes of their organs, etc.

For the theory of probability the advances of physical statistics and of a number of branches of technology raised a large number of absolutely new problems which did not fit into the confines of classical patterns. Physics and technology were interested in studying

processes, i. e., phenomena proceeding in time whereas the theory of probability had no general methods, no developed partial patterns for solving problems caused by the study of such phenomena.

There appeared an urgent need to develop a general theory of *stochastic processes* (of random variables *depending on one or more changing parameters*)[58]. The beginnings of such a general theory were due to the fundamental work of Soviet mathematicians, A. N. Kolmogorov and L. Ya. [A. Ya.] Khinchin. In a certain sense this theory has been developing the notions connected with sequences of dependent random variables introduced by Markov in the first decade of the 20th century (Markov chains). He only considered his theory as a mathematical discipline, but in the 1920s physicists converted it to become an effective tool for investigating nature.

Later, many scientists (S. N. Bernstein, V. I. Romanovsky, Kolmogorov, Hadamard, Fréchet, Doeblin, Doob, Feller and others) essentially contributed to the theory of Markov chains. Also in the 1920s, Kolmogorov, E. E. Slutsky, Khinchin and Lévy discovered a close connection between the theory of probability and the mathematical disciplines studying sets and the general notion of functions (set theory and the theory of functions of a real variable). Somewhat earlier Borel arrived at the same concepts. Their discovery proved extremely fruitful and it was in this direction that the final solution of the classical problems formulated by Chebyshev was found[59].

Lastly, we ought to indicate the work of Bernstein, Kolmogorov and Mises devoted to the construction of a logically harmonious theory of probability[60] capable to cover various pertinent problems formulated by natural sciences, technology and other branches of knowledge. However, in spite of considerable advances in the construction of a logical foundation of the theory of probability achieved by those authors, research in that direction is continuing intensively enough.

One of the reasons of this fact is the desire to understand the nature itself of randomness, to establish the connections between randomness of phenomena and their determinativeness. Nowadays, reassuring approaches to this great and important problem of general philosophical interest are discovered (if not its complete solution).

The further development of the theory of probability, just as each growing field of knowledge, requires an uninterrupted influx of fresh forces. It opens up a wide field for displaying the talents of young researchers, for their creative work. A deep interest in all sides of the theory of probability is needed for such talents to come into blossoming, an interest in the problems of its logical underpinning, in its connections with other mathematical disciplines, in disclosing new problems appearing in natural sciences (in physics, biology, chemistry etc), engineering, managerial work, economics and other areas of theoretical and practical activities.

Notes

1. The publishers listed the editions of this book (including those which had appeared in foreign languages). The seventh edition of 1970 was preceded by the sixth edition of 1964, the first to appear after Khinchin's death.

2, § 1.1. In the second example we should have rather mentioned *unsuccessful* results. However, successful in the theory of probability are the results which lead to the occurrence of the studied event. G&K.

3, § 1.1. This means that the particles are in an indifferent equilibrium. G&K.

4, § 1.1. In 1913, Markov (Petruszewycz 1983) studied the alteration of vowels and consonants in the Russian language. Knauer (1955) described early applications of statistics to linguistics.

5, § 1.3. Chebyshev (1845/1951, p. 29) and Boole (1851/1952, p. 251; 1854/2003, p. 246) defined the aim of the theory of probability as determining the probability of an event (of a proposition, as Boole suggested at first) by issuing from the given probabilities of other events. This definition seems to have persisted.

6, § 2.3. This is the principle of *mangelden Grunden* (Kries 1886, p. 6) which Keynes (1921/1973, p. 44) renamed *principle of indifference*. Laplace (1814/1995, p. 116) recommended to adopt hypotheses but rectify them *incessantly by new observations*.

7, § 2.3. Verification was necessary by studying all the drawings, but then only (any) one of them became sufficient.

8, § 3.1. A bulb is standard if it can burn for 1200 hours; otherwise, it is substandard. G&K.

9, § 3.1. This is easy to calculate. Of each 100 bulbs 700 in the mean are manufactured by the first factory; and of each of these 100 bulbs 83 are standard. Consequently, of the 700 bulbs $7 \cdot 83 = 581$ will be standard on the average. The other 189 standard bulbs are manufactured by the second factory. G&K.

10, § 3.3. This means that out of 100 specimens selected from the first skein 84 in the mean endure such a load and 16 do not. G&K.

11, § 3.3. Instead of a timely explanation of the pertinent principle, four significant digits are chosen instead of two! Same mistake in Example 1 below and in § 13.5.

12, § 4.3. Why should a location of a *destroyed* target be corrected? Same unimaginable attitude described in Example 1 below.

13, § 4.3. We *somehow* know the prior probabilities … This is the only remark (an oblique hint at that) about the serious shortcoming of the Bayes theorem.

14, § 4.3. A strangest idea. No one (at least until the advent of the computer) ever corrected or could have corrected artillery gunfire by the Bayes (or any other) theorem. For that matter, how many artillery men ever heard about Bayes?

15, § 4.3. Positive answer of a test is actually explained a few lines below.

16, § 5.1. In Example 1 of § 1.1 the figure *well known in demography* was 516. Below, the calculation is doubtful since different families apparently have differing inclinations to bear male (say) babies. In 1904, Newcomb (although certainly not a demographer) introduced three such classes of families (Sheynin 2002, pp. 153 – 154).

17, § 5.2. It is much more usual to say that those formulas describe the binomial distribution. The statement just below that $(n - k)$ is a large number is not generally true.

18, § 5.3. Owing to obvious difficulties, I omitted all the diagrams.

19, § 5.3. Actually, $197/17 \approx 11.6$.

20, § 6.1. Bernoulli discovered his theorem about 20 years before his death **[in 1705]** but it was only published in 1713. G&K.

The Bernoulli theorem is described unsatisfactorily. Bernoulli proved an extremely important existence theorem (and it was quite proper to say something about them) and studied the rapidity of the approach of the statistical probability to its theoretical counterpart. He did not yet know the (De Moivre −) Stirling formula and this study was therefore not satisfactory. In Chapter 4 of pt. 4 of his *Ars Conjectandi* Bernoulli formulated the inverse problem so that his theorem did not conform to his aim, but he alleged that he had solved both the direct and the inverse theorems. Only Bayes (Sheynin 2010) indicated that the inverse problem was less precise.

Contrary to the authors' statement, Bernoulli proved his existence theorem in an elementary way. They also failed to mention Poisson.

21, § 7.1. By definition, a meteorite is a small celestial body that *reached the Earth*.

22, § 7.1. This sentence is unfortunate.

23, § 7.2. It can be argued that no points should be awarded for missing the target. However, if a point means the right to shoot, even a miss provides a point. G&K.

24, § 7.2. The knowledge of the law of distribution of a random variable is indeed *sufficient*, but it is also the most possible knowledge.

25, § 7.2. We may also consider 2 as a possible value of $\xi + \eta$ having probability zero just like we did in table (I): for the sake of generality we stated that value 1 was possible. G&K.

And like stating that a probability equals 0 + 0.04.

26, § 8.1. The mean result is also random.

27, § 8.1. Beginning with De Moivre (1756, p. 3; possibly in the earlier editions of this book as well) expectation is simply defined rather than derived and the authors should have mentioned this fact. It is opportune to remark that Laplace (1812/1886, p. 189) had proposed the term *mathematical expectation* to distinguish it from the then topical *moral expectation*. His term is still being unnecessarily applied at least in French and Russian literature. *Statistical* probability (§ 1.1) is introduced as though it is the theoretical probability.

28, § 8.1. We assume that a part rejected when assembling a device is not used anymore. G&K.

An unsuccessful attempt therefore means that a part is lost, but the authors had not mentioned this circumstance.

29, § 8.1. An error of, say, ± 10 *m* means that both 10 and -10 *m* have probability 0.16. G&K.

30, § 8.1. *Always* is never stated in scientific definitions or statements, but the authors repeatedly (e. g., in § 9.1) apply this as also other unnecessary and possibly confusing words (*purely random*, in the beginning of § 7.1).

31, § 9.1. It would have been in order to say a few words about direct and inverse statements in general. Such statements are also mentioned in Note 20 and § 12.3.

32, § 10.1. On the mathematical meaning of *true value* see Sheynin (2007).

33, § 10.2. Is this a hint (repeated below) at empirical densities?

34, § 10.2.2. The authors did not introduce *standard deviation*.

35, § 10.2.3. *Shells fall around* … This is the only statement (and only an oblique hint) that the scatter of shells is two-dimensional.

36, § 10.3. Technologists decided that the creation of a theory of tolerances based on considerations and conclusions of probability theory was needed. G&K.

37, § 10.3. A strange example: a distance of 200 *m* measured so roughly! In § 11.2 the same distance is supposed to be measured 10,000 times!

38, § 11.1. The authors should have explained why the estimate (11.2) is *very rough*.

39, § 11.2. Poisson is forgotten once more (cf. Note 20). In § 11.3 Chebyshev is justly credited with a more general statement.

40, § 11.2. A few specimens each containing, say, 100 – 200 *g* are selected, whereas the entire amount of wheat measures tens and perhaps hundreds of tons of grain. G&K.

A few words about sampling in general would have been in order.

41, § 11.3. A wrong statement. Systematic errors are unavoidable and there always exists some dependence between observations. It was understood long ago that an excessive number of observations is useless. See Sheynin (1996, pp. 97 – 98).

42, § 11.4. Fechner (Sheynin 2004, pp. 60 – 61) discussed the origin of the errors of reading and their influence, but hardly satisfactorily. Geodesists never considered the errors of reading separately from all other errors. Cournot (1843, § 139), certainly not a practitioner, thought otherwise and his considerations are properly forgotten and I doubt that the authors could have justified their statement.

Moreover, contrary to their statement, the error of reading is not constant and the error of the arithmetic mean of readings (of two or three at most) is not the same as the error of one reading.

Mathematicians are generally ignorant of the theory of errors. In the beginning and mid-19th century French scientists including Poisson had been enraged by

Legendre's alleged mistreatment at the hands of Gauss, and to their own disadvantage did not read that great scholar. Laplace knew better than that but he kept to his own almost useless theory of errors nevertheless venerated for many decades. And even Chebyshev (who included the theory of errors in his lectures) did not study Gauss. See Sheynin (1996).

43, § 12.1. A few lines above the law of distribution was assumed known.

44, § 12.1. The scatter of shells is also discussed in Example 5 of § 12.4 and in the *Conclusions* and each time somewhat differently. Factors influencing crop capacity are somewhat differently mentioned in §§ 7.1 and 10.1.

45, § 12.1. Cf. also the *Conclusions*. G&K.

46, § 12.2. How a deviation (or an error) of a few metres can be important when measuring a distance between settlements?

47, § 12.3. For readers acquainted with elements of higher mathematics we note that the equation of the curve representing a normal law is

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

Here, $\exp(x) = e^x$; $e = \ldots$ is the base of natural logarithms; $\pi = \ldots$ is … and a and σ^2 are the mean value and variance of the random variable. The knowledge of the analytical form of the normal law can considerably simplify the acquaintance with the following text, which is however easily understood to readers unacquainted with higher mathematics as well. G&K.

Why the notion of curves of distribution (§ 12.2) is not similarly explained?

48, § 12.3. Symbol $\sum_k$ should be understood as $\sum_{k=1}^{n}$. G&K.

The authors had not explained the latter symbol although on p. 25 did explain the meaning of three dots (omitted here). This is a usual occurrence: authors of popular writings begin explaining everything but soon have to abandon this intention.

49, § 12.4. In addition to Note 44 I remark that the error of aiming an artillery gun certainly changes from shot to shot.

50, § 13.1. This notation is at variance with the previous notation $v_x(t)$ etc.

51, § 13.2. A few words should have been added about chaotic motion.

52, § 13.3. Capacities of locks etc are mentioned under consumer services!

53, § 13.3. The problem is at least heuristically connected with the central limit theorem which is only mentioned in the *Conclusions*.

54, Conclusions. Translated by David (1962, p. 115).

55. The theory of probability came to a dead end because Laplace forcefully transferred it from pure mathematics to an applied science. For many decades the splendid work of Chebyshev and his students had barely interested mathematicians because of that very circumstance, witness Markov's report of 1921 (Sheynin 2006, p. 152): *The theory of probability was usually considered as an applied science in which mathematical rigour was unnecessary*. The renewal of the situation began with Lévy.

56. In general, the acting factors should be expressed by differing random variables.

57. We may indeed expect normality, but not at all always.

58. At the end of § 13.1 and the beginning of § 13.2 only one parameter was mentioned.

59. I doubt that such problems existed.

60. Kolmogorov published several pertinent contributions of which we mention the lesser known note (1983). Bernstein (1917) seems to have been largely ignored; Khinchin (1961) published an essay on the Mises theory. Uspensky et al (1990, § 1.3.4) stated about that theory: *Until now, it proved impossible to embody Mises' intention in a definition of randomness that was satisfactory from any point of view*.

Bibliography

Abbreviation: **S, G,** No. … = my website sheynin.de copied by Google: Oscar Sheynin, Home. Downloadable Document No. …

Anonymous (1955, in Russian), On the role of the law of large numbers in statistics. *Uchenye Zapiski po Statistike*, vol. 1, pp. 153 – 165.

Bernstein S. N. (1917, in Russian), An essay on an axiomatic justification of the theory of probability. *Sobranie Sochinenii* (Coll. Works), vol. 4. Moscow, 1964, pp. 10 – 60. Translation: S, G, in No. 6.

Boole G. (1851), On the theory of probability. In author's *Studies in Logic and Probability*, vol. 1. London, 1952, pp. 247 – 259.

--- (1854), *Laws of Thought*. Amherst, N. Y., 2003.

Chebyshev P. L. (1845). Opyt elementarnogo analiza teorii veroiatostei. *Polnoe Sobranie Sochinenii* (Complete Works), vol. 5. Moscow – Leningrad, 1951, pp. 26 – 87.

Cournot A. A. (1843), *Exposition de la théorie des chances et des probabilités*. Paris, 1984. Editor, B. Bru. Translation: **S, G,** No. 54.

David F. N. (1962), *Games, Gods and Gambling*. London.

De Moivre A. (1718), *Doctrine of Chances*. London, 1738, 1756. New York, 1967.

Gnedenko B. V. (2001), *Ocherk Istorii Teorii Veroiatnostei* (Essay on the History of the Theory of Probability). Moscow, 2009.

Gnedenko B. V., Sheynin O. (1978, in Russian), Theory of probability. A chapter in *Mathematics of the 19th Century*, vol. 1. Editors, A. N. Kolmogorov, A. P. Yushkevich. Basel, 1992, 2001, pp. 211 – 288.

Keynes J. M. (1921), *Treatise on probability*. *Coll. Writings*, vol. 8. London, 1973.

Khinchin A. Ya. (1937, in Russian), The theory of probability in pre-revolutionary Russia and in the Soviet Union. *Front Nauki i Tekniki*, No. 7, pp. 36 – 46. Translation: **S, G**, in No. 7.

--- (1943, in Russian), *Mathematical Foundations of Statistical Mechanics*. New York, 1949.

--- (1948, in Russian), *Eight Lectures on Mathematical Analysis*. Heath & Co., 1965.

--- (1953), *Kratkii Kurs Matematicheskogo Analiza* (Brief Course on Mathematical Analysis). Moscow.

--- (1961, in Russian), The Mises frequency theory and modern ideas of the theory of probability. *Science in Context*, vol. 17, 2004, pp. 391 – 422.

Knauer K. (1955), Grundlagen einer mathematischen Stilistik. *Forschungen u. Fortschritte*, Bd. 29, pp. 140 – 149.

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

--- (1983), On logical foundations of the theory of probability. *Lecture Notes Math.*, No. 1021, pp. 1 – 5.

--- (1955, in Russian), [Description of his report at a statistical conference of 1954]. Anonymous (1955, pp. 156 – 158). Translation in **S, G,** No. 6.

Kries J. von (1886), *Principien der Wahrscheinlichkeitsrechnung*. Tübingen, 1927.

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

--- (1814, in French), *Philosophical Essay on Probabilities*. New York, 1995.

Novikov S. P. (2002, in Russian), The second half of the 20th century and its result: the crisis of the physical and mathematical community in Russia and the West. *Istoriko-Matematich. Issledovania*, vol. 7(42), pp. 326 – 356.

Petruszewycz M. (1983), Description statistique de textes littéraires russes par la méthode de Markov. *Rev. Etudes Slaves*, t. 55, pp. 105 – 113.

Sheynin O. (1996), *The History of the Theory of Errors*. Egelsbach.

--- (1998), Statistics in the Soviet epoch. *Jahrbücher Nat.-Ökon. und Statistik*, Bd. 217, pp. 529 – 549.

--- (2002), Newcomb as a statistician. *Hist. Scientiarum*, vol. 12, pp. 142 – 167.

--- (2004), Fechner as a statistician. *Brit. J. Math. & Stat. Psychology*, vol. 57, pp. 53 – 72.

--- (2006, in Russian), On the relations between Chebyshev and Markov. *Istoriko-Matematich. Issledovania*, vol. 11 (46), pp. 148 – 157.

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

--- (2010), The inverse law of large numbers. *Math. Scientist*, vol. 35, pp. 132 – 133.

Uspensky V. A., Semenov A. L., Shen A. Kh. (1990, in Russian), Can an (individual) sequence of zeros and ones be random. *Uspekhi Matematich. Nauk*, vol. 45, pp. 105 – 162. This journal is being translated from cover to cover as *Russ. Math. Surveys*.

Yaglom A. M., Yaglom I. M. (1957, 1960, 2007, in Russian), *Probability and Information*. Kluwer, 1983.