Studies in the History of Statistics and Probability

vol. 5

Collected Translations

E. S. Ventzel, O. Sheynin, L. Z. Rumshitsky

Elementary Probability

Compiled and translated by Oscar Sheynin

Berlin
2014

Annotation

The three contributions translated below are very different. The first is very elementary and as such deserves some attention as being perhaps methodically unique. The third booklet went at least through five editions: 1960, 1963, 1966 (here translated), and 1970, 1976 which I have regrettably not seen. This circumstance all by itself is remarkable. My own contribution was an attempt to link a rather elementary exposition with the history of probability. I accompanied the first and the third contributions by notes, and certainly do not repeat here their essence. There certainly exist good more or less elementary expositions of probability published in English. Some of them are referred to below and I can also mention

F. Mosteller., R. E. K. Rourke, G. B. Thomas (1961), *Probability and Statistics*. New York, 1965.

F. Mosteller (1965), *Fifty Challenging Problems in Probability*. New York.

A few words about Rumshitsky's booklet. He had barely touched on statistics and, as I see it, italicised too many sentences and passages; in this respect, I did not always follow him. Then, a large portion of his text was in small print which was not necessary to preserve.

Contents

E. S. [Elena Sergeevna] Ventzel

The Theory of Probability (the First Steps)

Е. С. Вентцель, *Теория вероятностей (первые шаги).* Москва, 1977

Contents

Chapter 1. What about Is the Theory of Probability?

The theory of probability occupies a special place in the family of mathematical sciences. It studies the laws of a special type governing random experiments[1]. Later, if and when specifically studying this theory, you will gain a profound knowledge of such laws, but my aim is far more modest. I am attempting to acquaint the reader with some elementary notions about the theory of probability, its problems and methods, its possibilities and boundaries.

Nowadays the development of science is characterized by a general mighty offensive of stochastic (statistical) methods on a wide front, an attack on every branch of knowledge. Today, each engineer, researcher and manager ought to be informed about the elements of theory of probability. However, experience shows us that, generally speaking, beginners run into difficulties when studying it. For a person accustomed to quite different traditional scientific ways it is not easy to become accommodated to its specific features. Most difficult are usually the first steps towards understanding and applying stochastic methods. The sooner this peculiar psychological barrier is removed, the better will it be.

In my booklet, without aspiring to describe systematically the theory of probability, I am attempting to help the beginner with these same first steps. And, in spite of the free and easy (at times, even comic) form of exposition, an attentive reader will on occasion have to think hard.

First, only a few words about *random events*. Suppose that the outcome of some experiment (or *trial*) cannot be predicted. For example, we toss a coin and cannot say whether heads or tails will appear. Or, we blindly draw a card from a pack and cannot predict its suit. Another example: we arrive at a bus station regardless of the bus timetable; how long will we have to wait? As chance would have it! Finally, how many articles manufactured under certain conditions will be defective?

All these examples describe random events with unpredictable outcomes. When experiments with indefinite outcomes are being repeated, their results will change. Thus, a precise balance will generally show differing values of the weight of an object. Why these differences? The conditions of the experiments appear identical, but the outcome of each is influenced by many small and hardly revealable factors causing in total an indefinite outcome.

And so, suppose that some experiment whose outcome is not known beforehand is random. Each fact that can occur or not is called a *random event* (or simply an *event*)[2]. Thus, a coin toss can lead or not to an event *A*, to the appearance of heads. Another experiment, a toss of two coins, can lead or not to an event B, the occurrence of two heads. One more example. A lotto consists of 49 numbers 6 of which are drawn. Each of the following happy events can occur: A, B and C, denoting 3, 4 or 5 guessed numbers. Also, the happiest (but the most unlikely) event D, denoting all 6 guessed numbers[3].

The theory of probability allows us to measure the *likelihood* (or probability)[4] of various events, to compare them according to their probabilities, and, what is the most important, thus to *predict* the outcomes of random phenomena ! You will possibly by angered here by understanding nothing at all. Just above you read that random phenomena are unpredictable, but now, predictions!

Just a minute, be patient. We can only predict random events having a high likelihood or probability. And it is the theory of probability that allows us to determine which events belong to that class. Let us discuss probabilities of events. Obviously, not all random events are equally probable, they can be more or less probable.

Which outcome of a roll of a die do you think is more probable: A, 6 points; or B, an even number of points? A difficult problem? Then shut my booklet and forget about it. But no! This is entirely unlikely; on the contrary, you will say at once: What kind of a problem is this? Obviously, event B. And you will be in the right since an elementary understanding of the notion *probability of an event* is a distinctive feature of each human being not lacking in common sense.

We are surrounded by random phenomena and random events, from childhood we are accustomed to estimate somehow those probabilities when contemplating our actions, to separate them into probable, unlikely, and hardly feasible. When the probability of an event is very low, common sense tells us not to reckon seriously on its appearance. Suppose a formula interesting us is placed somewhere in a book 500 pages long. Can we seriously expect to find it by blindly opening the book? Apparently not. Such an event is possible but unlikely.

Now let us agree how to calculate (to estimate) the probability of random events. First of all, we should assign a probability for a *certain event*, i. e., an event that will certainly occur in a given experiment. For example, it is certain that the number of points achieved in a roll of a die will not exceed 6. By definition, the probability of a certain event is 1. And probability 0 is assigned to an *impossible event*, i. e., an event that cannot occur in a given experiment at all. Thus, a negative number of points cannot occur in a roll of a die.

Denote the probability of event A by $P(A)$, then obviously

$$0 \leq P(A) \leq 1. \qquad (1.1)$$

Keep in mind this most important property of probability! If, while solving a problem, you get a probability higher than 1 (or, still worse, a negative probability), be sure that you have been mistaken. One of my students failed to understand this. Arriving at $P(A) = 4$, he stated that the event was *more than probable*[5].

But to return to our considerations. I repeat that the probability of an impossible event is 0 and that a certain event has probability 1. Then, the probability of any random event A is a number contained between 0 and 1. It shows the *part* of the probability of a certain event possessed by that given event. Later, you will learn how to calculate the probabilities of random events when dealing with some simple problem. Now, however, we have to think about certain points of principle connected with the theory of probability and its applications.

And, first of all, why should we be able to calculate probabilities? It is certainly sufficiently interesting in itself to know how to estimate numerically the degrees of likelihood of various events, to compare them one with another.

But our final aim is different: we wish to *predict* the outcomes of experiments concerning random phenomena by issuing from their calculated probabilities. Indeed, there exist such experiments whose outcomes are predictable in spite of randomness, predictable either exactly or approximately, with certainty or, so to say, practically certainly. To reveal a special class of events *practically certain* and *practically impossible*, to discern events *A* for which $P(A) \approx 1$ and 0, is one of the most important problems of the theory of probability.

Suppose that each of 100 men tosses his own coin. Event *A* is the appearance of 100 heads. It is theoretically possible to imagine such a freak of chance, but its probability is negligible; we will soon find out that it equals $1/2^{100}$. Event *A* can be considered practically impossible while the contrary event $\bar{A}$ meaning that there occurs at least 1 tails is practically certain. In problems concerning probabilities practically certain and practically impossible events always occur in pairs, just as above.

If our calculations show that some event *A* is practically certain, we may *predict* its occurrence, although not for sure, but almost so. Not a little success when dealing with random phenomena! We may thus almost for sure predict the maximal error in our computer calculations; the maximal and minimal yearly number of spare parts needed by a garage; the maximal and minimal number of successive shots when shooting at a target; the maximal [relative] number of defective articles.

Note that such predictions are generally possible when studying *many* homogeneous random events rather than a single isolated event. It is impossible to say beforehand whether the outcome of a coin toss will be heads or tails, and no theory of probability whatever helps here. However, we can predict the boundaries within which the number of heads will be contained after, say, 500 or 1000 tosses. Such examples will be offered below; predictions are formulated not for sure but almost so, and are realized not without exception, but in most cases.

People are often asking: how high should be the probability of an event for considering it practically certain? Should it be, say, 0.99 or 0.995, or even 0.999? No definite answer is possible. All depends on how important is the success of the prediction, and what will threaten us if it fails? Suppose we predict that, with probability 0.99, if travelling by a certain type of public transportation, we will not be late for work more than by 10 minutes. Can we consider this event practically certain? Yes, as it seems. The same question about a favourable landing of a spaceship? Obviously, probability 0.99 is not here enough!

Now, keep in mind that any prediction provided by the theory of probability is always characterized by two features:

1. It is not offered for sure, but *almost for sure*, i. e., with a high probability.

2. The researcher himself assigns the numerical value of this probability (of this *confidence level*) more or less arbitrarily, but in

accordance with common sense and allowing for the importance of the success of the prediction.

If, after all these reservations, you are not yet definitively disappointed in the theory of probability, go on reading and get acquainted with some elementary methods of reckoning the probabilities of random events. Some idea about these methods you already apparently possess. Answer, for example, this question: What is the probability of heads in a coin toss? Almost surely (with a very high probability) you will say at once: 1/2. And this is the correct answer provided that the coin is symmetric and of regular shape[6], and the outcome *edgeways* is considered practically impossible.

Just as easy is the question about the appearance of 6 points in a roll of a die. You will almost surely answer, 1/6 (with the same reservations allowed for). How did you arrive at this answer? Apparently, you have noted that there are six possible outcomes and that, owing to the symmetry of the die, they are all equally probable. It was therefore natural to assign probability 1/6 to each, which is what you did quite correctly.

But now, what is here the probability of the occurrence of more than 4 points? You will probably answer, again correctly, 1/3. Indeed, two equally probable outcomes, 5 and 6 points, are *favourable* for the event. You divided 2 by 6 and obtained the correct answer. Bravo! Without suspecting it, *you have applied the classical method of calculating probabilities according to the pattern of chances*.

But what is that, the pattern of chances? First, we introduce a few terms; in the theory of probability, just as in many other sciences, terminology is important. Suppose that an experiment has possible outcomes A_1, A_2, …, A_n. Events A_1, A_2, …, A_n are called *incompatible* if they mutually exclude one another, if no two of them can occur at the same time.

They form a *complete group* if they exhaust all possible outcomes, if the non-appearance of all of them is impossible. They are *equally probable* if the conditions of the experiment ensure an equal possibility (probability) of the appearance of each of them. If those outcomes possess all the three properties (are incompatible, form a complete group and are equally probable) they are called *chances* and the experiment is said to be reduced to the *pattern of chances*[7].

The experiment of coin tossing is thus reduced to the pattern of chances since the two outcomes, A_1 and A_2, possess all the three abovementioned properties. The same can be stated about the rolls of a die with six possible outcomes.

Consider now the toss of two coins. If thoughtless and hasty, you will be quick to mention three events: B_1, two heads; B_2, two tails; B_3, heads and tails. You will be wrong! These events are not *chances* since they are not equally probable; B_3 is twice as probable as each of the other two. The *real* chances are A_1 and A_2, as also are B_1 and B_2, and A_3 and A_4: heads on the first (on the second) coin, tails on the second (on the first) coin.

In our next example we will for the first time use the traditional *urn* with balls, or, simply speaking, a container with some number of balls of various colours. They are thoroughly shuffled and do not differ to the touch which ensures an equal probability of drawing any of them. These conditions will be implied in each problem connected with urns.

Each problem in which the experiment is reduced to the pattern of chances can be considered as an urn problem. These latter problems constitute a single language of sorts capable of describing instances most variable in appearance. So let us have an urn with 7 balls, 3 white and 4 black. A ball is blindly drawn and it is required to enumerate the appropriate chances. Here, it is once more possible to name thoughtlessly two events, B_1 and B_2, the occurrences of a white and a black ball. If you were thus tempted, you are not fit for the theory of probability. However, you will quite likely reject such an answer, you have already understood that B_1 and B_2 are not equally probable. Here, we have 7 rather than 2 cases, as many as there are balls. There cases are incompatible, they form a complete group and are equally probable and they therefore represent chances.

Now, is it possible to form a group of chances for each experiment? No, far from it. If, for example, the experiment consists in tossing an irregular (a bent) coin, the appearances of heads and tails will not anymore be chances since they are not equally probable. It is even possible to bend a coin in such a way, that one of these outcomes will become impossible. The toss of an irregular coin cannot be reduced to the pattern of chances. For such a reduction the experiment should possess some symmetry and thus to enable equal probabilities of the outcomes.

That symmetry is sometimes achieved by physical symmetry (of the coin or die) or by shuffling the elements involved which ensures an equal probability of the extraction of any one of them. Most often such symmetry is observed in artificially arranged experiments if only special measures are implemented

Typical examples are provided by games of chance. Note that the development of the theory of probability began with their analysis. If an experiment is reduced to the pattern of chances, the probability of any event A included there can be calculated as *the ratio of the number of chances favourable for A* (m_A) *to the total number of chances* (n)

$$P(A) = m_A/n. \qquad (1.2)$$

This is the so-called *classical formula*. It has been applied from the very origin of the science of random phenomena. For a long time it was even considered as the *definition of probability*. Experiments which did not possess symmetry had been artificially adjusted to fit the pattern of chances[8]. In our time, probability, as well as the methods of discussing its theory, is regarded from another point of view.

For us, formula (1.2) is not universal, although it ensures calculations of the probabilities of events in some simplest experiments. In the subsequent chapters you will see how to calculate the probabilities of events when the appropriate experiment is not reducible to the pattern of chances.

We will now consider a number of problems for illustrating the calculation of probabilities of random events according to formula (1.2). Some of them are very easy, the other ones are not.

Ex. 1. Two coins are tossed. Required is the probability of the appearance of at least one heads (of event A).

Here, we have 4 cases (as explained above), three of them favourable for A, so $P(A) = 3/4$.

Ex. 2. An urn contains 3 white and 4 black balls. A ball is drawn, and required is the probability that it is white (event A).

Here, $n = 7$, $m_A = 3$, $P(A) = 3/7$.

Ex. 3. A ball is drawn from the same urn and put elsewhere without noting its colour. A second ball is drawn. Required is the probability that it is white (event A).

Pondering about these circumstances, we [apparently] convince ourselves that the preliminary extraction does not influence event A. It remains as it was in Ex. 2 (3/7). However, before the second ball was extracted, the urn had contained 6 rather than 7 balls, so was the number of cases really 7? Unless and until we know the colour of the first ball, the number of cases remains as it was, 7. To convince ourselves, we will preliminarily draw 6 balls rather than 1 again without noting their colour.

The probability that the only one left is white will still be 3/7 since it is irrelevant whether we draw it or leave it in the urn. You are still doubtful? Then imagine a dark room. We draw all the 7 balls, throw 2 of them somewhere on the floor and put the rest on a cupboard. Then we accidentally step on one of those two balls and require the probability that it is white. Are you still in doubt? Well, nothing will help you since all our arguments are exhausted[9].

Ex. 4. Two balls are drawn at the same time from the same urn. Required is the probability that both are white (event A).

This problem is somewhat more difficult since it is not so easy to calculate n and m_A in formula (1.2). We will have to decide in how many ways we can draw these two balls, and to draw both white balls. Such problems belong to the subject of a special science, combinatorics, a branch of elementary algebra. Here, we only need one of its formulas for calculating the number of combinations of k things taken s at a time and differing by their composition but not by the order of their elements:

$$\left(\frac{k}{s}\right) = \frac{k(k-1)...(k-s+1)}{s!}, \quad \left(\frac{k}{s}\right) = \left(\frac{k}{k-s}\right). \qquad (1.3, 1.4)$$

In our example, $k = 7$, $s = 2$ and $n = 21$. Now we have to find the number of ways for selecting 2 out of the 3 white balls. Here, $k = 3$ and $s = 2$, so $m_A = 3$. Finally, $P(A) = 3/21 = 1/7$.

Ex. 5. Three balls are drawn at once from the same urn (put up with it for a while longer, we will soon leave it). Required is the probability that 2 of them will be black, and 1, white (event A).

Here, $k = 7$, $s = 3$ and $n = 35$. Now, two out of 4 black balls can be selected in 6 ways ($k = 4$ and $s = 2$) and each such combination should go with each combination of the selected white balls. The total number of favourable cases will be $6 \cdot 3 = 18$, and $P(A) = 18/35$.

We are now prepared to solve the following general problem.

Problem. An urn contains *a* white and *b* black balls and *k* balls are drawn. Required is the probability that among these there will be *l* white balls (and therefore *k* – *l* black balls); $l \le a$, $k - l \le b$.

The number of cases is

$$n = \left(\frac{a+b}{k}\right).$$

Now, we select *l* white balls from *a* and *k* – *l* black balls from *b*, then take each of the first combinations with each of the second ones so that

$$P(A) = \left(\frac{a}{l}\right)\left(\frac{b}{k-l}\right) \div \left(\frac{a+b}{k}\right). \qquad (1.5)$$

This formula can be applied in various settings, for example when solving problems concerned with sampling acceptance. An urn is then replaced by a batch containing defective (black balls) and quality articles (white balls) with *k* balls in a trial sample. One more example which will perhaps interest you.

Ex. 6. A gambler selected 6 numbers of a lottery containing 49 numbers. Required is the probability that he had guessed 3 numbers out of the drawn 6.

But this was explained in the previous example. Just imagine an urn with 6 white and 43 black balls. Required is the probability that out of the drawn 6 balls 3 will be white. According to formula (1.5) with a = 6, b = 43, k =6 and l = 3,

$$P(A) = \left(\frac{6}{3}\right)\left(\frac{43}{3}\right) \div \left(\frac{49}{6}\right) = \frac{6 \cdot 5 \cdot 4 \cdot 43 \cdot 42 \cdot 41 \cdot 4 \cdot 5}{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}.$$

If not feeling lazy, calculate this result. The probability is very low, ca. 0.0176 or only 1.8%. The probabilities of guessing 4 or 5 or all 6 numbers (a wonder!) is still lower (much lower).Try and calculate them if you are so especially inclined.

Chapter 2. Probability and Relative Frequency

You have acquainted yourself with the subject of the theory of probability, with some of its main notions and with the calculation of probabilities of events according to the so-called classical formula (1.2). It does not follow, however, that you are well equipped for practically applying the theory of probability. The sphere of application of that formula is regrettably not as vast as desired. It is only useful for experiments symmetrical with respect to symmetry of possible outcomes (reducible to the pattern of chances). Such experiments mostly belong to games of chance in which symmetry is ensured by special measures[10]. Unlike former times, nowadays professional gamblers are not really numerous, and the practical importance of formula (1.2) is very restricted. So how should we deal with the probabilities of events in most cases? Do they exist? If they do, how to calculate them?

Here, we ought to introduce a new main notion of the theory of probability, the notion of *relative frequency of an event*. Let us approach it from some distance. Suppose we throw an irregular asymmetric die and let event A be the occurrence of 6 points. Formula (1.2) is inapplicable here[11] and we cannot say that $P(A) = 1/6$. So is it higher or lower than that value? And how to find this probability, at least approximately?

Any reasonable man will say: let us roll the die many times and see how often (relatively) the event occurs. This frequency can be assumed as the probability of A. What can you say? Our reasonable man is certainly in the right. Without knowing it, he applied the notion of frequency which we will now rigorously define.

Frequency of an event in a series of experiments is the ratio of the number of those of them in which that event had occurred to the total number of experiments.

This frequency is also called *statistical probability*. Statistics of mass random phenomena is the foundation for determining the probabilities of events with possible asymmetrical outcomes. We will denote frequency by P^* so that

$$P^*(A) = M_A/N. \qquad (2.1)$$

Here, N is the total number of experiments and M_A, the number of those of them in which the event A had occurred. Formulas (1.2) and (2.1) are similar in appearance but absolutely different in essence. The former *theoretically* computes the probabilities of events given the conditions of an experiment, whereas the latter *experimentally* determines the frequency of the events. For applying it, we need experimental, statistical data.

Let us think awhile about the essence of frequency of an event. It is quite obvious that some connection exists between it and probability. More probable events generally occur more often than less probable ones, but the two notions are not at all identical. The similarity, the *kinship* between them becomes the more noticeable the more trials are made. With a small number of them frequency is essentially random, can considerably deviate from probability.

For example, in 10 coin tosses heads can indeed appear 3 times, its frequency is then 0.3, very different from its probability, 0.5. But when the number of experiments increases, frequency gradually loses its random essence[12]. Random circumstances accompanying each experiment compensate each other and frequency generally stabilizes and with slight fluctuations approaches some mean constant magnitude. It is natural to suppose that that magnitude is nothing but the probability of the event.

We can verify that statement, although obviously only for those events whose probabilities can be calculated by formula (1.2), i. e., for those experiments which are reducible to the pattern of chances. And that statement proved to be correct. You can check it yourself by choosing a simple example, by coin tossing, let us say. The frequency of heads will approach its probability, 0.5. Toss the coin 10, 20, … times (until patience lasts) and calculate that frequency. For sparing efforts and time apply a simple ruse, throw coins by the dozen (but certainly shake them thoroughly beforehand). […]

You will not be the first to conduct such experiments, eminent scholars did not shrink from them. Thus, Pearson[13], a celebrated statistician, made 24 thousand coin tosses and got 12,012 heads so that its frequency was very near to 0.5. Many experiments were made with a die and the frequency of the occurrence of its faces proved to be near 1/6. The approach of the frequency to probability can be considered experimentally proven.

Stability of the frequency given a large number of homogeneous experiments is one of the most typical regularities observed in mass random phenomena. When repeating the same experiment many times (and ensuring their *independence*) the frequency of the studied event becomes ever less random[14], more equable and approaches a constant. For experiments reducible to the pattern of chances it is possible directly to convince ourselves that that constant is nothing but the *probability of the event*. Suppose, however, that the experiment is not reducible to that pattern, but that the frequency becomes stable and approaches a constant. Well, we will then assume that the formulated rule is holding and call the approached constant the *probability of the event*.

We have thus introduced probabilities not only for events and experiments reducible to the pattern of chances, but for other events and experiments as well, if only *the stability of the frequency* persists. But what can be said about that stability? Does it exist for all random phenomena? Not for all, but for many. Let us explain this not really simple statement. Try to consider carefully this explanation since it will save you from possible mistakes.

When discussing stability, we assumed that the same experiment (coin toss, a roll of a die) can be repeated indefinitely. Indeed, nothing prevents us from repeating *such an experiment* any number of times provided we have the necessary time. But sometimes not we ourselves, but nature *arranges* experiments, and we only observe their results. In such cases stability of the frequency cannot be guaranteed beforehand, it ought to be verified.

Suppose that an *experiment* is a male birth, and that we are interested in its probability. Nature arranges it a great many times yearly. Is the frequency of such random phenomena stable? Yes, as experimentally established, it is very stable. It barely depends on the geographic location of the country in question, on the nationality or age of the baby's parents etc. It is somewhat higher than 0.5 (approximately equals 0.51)[15].

Frequencies are stable (at least not over very large time intervals) is present in such random phenomena like, for example, failures of technical equipment, appearance of defective articles, wrong work of machinery, morbidity and mortality of a population, meteorological and many biological phenomena. It is this stability that allows us to apply successfully stochastic methods for studying those phenomena, predicting and governing them[16].

There also exist such random phenomena for which that stability is doubtful or does not even exist. In such cases it is meaningless to mention a long number of *homogeneous* experiments, since a sufficiently large ensemble of statistical data does not exist (or cannot in principle be obtained). Such phenomena can include some events which seem to be more or less plausible, but no definite probability can be assigned them. Thus, it is hardly possible (and hardly expedient) to calculate the probability that in three years women will wear long skirts (or men grow moustaches). There is no appropriate ensemble of statistical data and successive years, if considered as experiments, cannot be thought to be in any sense homogeneous.

Another example (a question) in which a probability of an event is even less sensible: What is the probability that organic life exists on Mars? Whether such life exists there will apparently be found out in a few years[17]; many scientists consider it quite plausible. But some degree of plausibility is not yet probability. When guesstimating probability, we inevitably find ourselves in a world of vague fantasies; when wishing to deal with genuine probabilities, we should base ourselves on sufficiently vast statistics. But in the case above this latter condition is certainly lacking: Mars is unique!

So let us specify: we will only discuss probabilities of events in experiments not reducible to the pattern of chances when they belong to the class of mass random phenomena with stable frequencies. Whether these are stable or not is usually decided by common sense. Is it possible to repeat sufficiently many times an experiment without essentially changing its circumstances? Can we hope to collect the necessary statistical data? If the researcher intends to apply stochastic methods, he himself ought to answer these questions.

To dwell now on yet another question directly connected with the previous discussion. When speaking about the probability of an event in some experiment it is necessary first of all to list carefully the main conditions of that experiment. They are supposed to be fixed rather than changeable once the experiment is repeated. One of the most common mistakes in practical applications of the theory of probability (especially made by the beginners) consists of speaking about the probability of an event without specifying the conditions of the appropriate experiment or the statistical ensemble of random phenomena in which that probability could have revealed itself as frequency.

Thus, it is utterly meaningless to speak about the probability of such an event as the delay of a train. Of what train? Freight or passenger train? Wherefrom and where does it go? Along which railway line? Only after specifying all these circumstances we are allowed to consider the probability of that event as a definite number. We are thus warning you about *reefs* threatening those who are not only interested in amusing petty problems about coins, dice and cards[18], but desire to apply stochastic methods for attaining veritable aims.

Suppose now that all those conditions are met: we can conduct sufficiently many homogeneous experiments, and the frequencies are stable. Then, having a long series of experiments, we may approximately equate the frequency of an event to its probability.

Indeed, we have agreed that that frequency approaches probability as the number of experiments increases. This seems to be, but is not really very simple. The interrelation of frequency and probability is rather delicate.

Let us think for some time about the term *approaches*. What does it mean? What a strange question, you will possibly think. *Approaches* means *comes ever nearer*. What is there to think about? There is something indeed. We are dealing with random phenomena, with everything happening in a special way, out of the ordinary.

We know that the sum of the geometric progression

$$1 + 1/2 + 1/2^2 + \ldots + 1/2^n$$

indefinitely approaches 2 as n increases. The more terms we take, the nearer will their sum be to their limit, which is absolutely certain. However, we are unable to make such categorical statements when dealing with random phenomena.

Yes, frequency generally approaches probability, but in its own way: not exactly for sure, but, with high probability, *almost so*. It can happen that even after a very long series of experiments frequency will essentially deviate from probability. The probability of that occurrence is very low, and the lower, the larger is the number of the experiments.

Suppose we toss a coin 100 times. Can it happen that the frequency of heads will essentially differ from its probability, 0.5? Can it be 0 (no heads at all)? Such an outcome is theoretically possible (it does not contradict any law of nature), but its probability is very, very low. Let us calculate it; happily, we are already able to solve such simple problems. First of all, calculate the general number n of cases. There are two outcomes for each coin, and each outcome can go with any outcome of any of the other coins, so that $n = 2^{100}$. Then, there is only one favourable case, and its probability is therefore $1/2^{100}$. This number has 30 zeros after the decimal point and we may safely regard an event having such probabilities practically impossible. Actually, even lesser deviations of the frequency from probability are also practically impossible.

So how large are the practically possible deviations given a long number N of experiments? I will write down the formula providing the answer to that question. Regrettably, I am not in position to prove it although some justification is given below. At present, you can only trust me[19]. Suppose that event A arrives with probability p in each of N experiments. Then with probability (*confidence level*) 0.95 the frequency $P^*(A)$ of that event will be contained within the interval (*the confidence interval*)

$$p \pm 2\sqrt{p(1-p)/N}. \qquad (2.2)$$

This means that almost always (or, more precisely, in 95% of all cases) the frequency will not be beyond that interval. Yes, in 5% of the cases we will be wrong, but *nothing ventured, nothing gained*. Or, when fearing mistakes, do not predict random phenomena since your statements will come true not *for sure*, but only *almost so*.

A quite another point is that someone will decide that probability 0.05 of a mistake is too high. Then we can play safe and apply a somewhat wider confidence interval

$$p \pm 3\sqrt{p(1-p)/N}. \qquad (2.3)$$

It corresponds to a very high confidence level of 0.997. But suppose that we demand a complete certainty of prediction? Then we will only be able to say that the frequency will not go beyond interval [0, 1] which is a rather trivial and obvious statement.

In Chapter 1 I had stated that the probability of a practically certain event (of the confidence level) is assigned to some degree arbitrarily. Let us agree that, when estimating the precision of determining probability by frequency we will be satisfied by a modest confidence level of 0.95 and apply formula (2.2). After all, nothing disastrous happens if we will sometimes be mistaken.

And so, suppose we toss a coin $N = 100$ times. We have $p = 1 - p = 0.5$ and formula (2.2) provides

$$0.5 \pm 2\sqrt{0.25/100} = 0.5 \pm 0.1.$$

With probability (confidence level) 0.95 we can thus predict that in 100 tosses the frequency of heads will deviate from its probability not more than by 0.1. Well … the deviation, to put it bluntly, is considerable. How can we decrease it? Apparently, by increasing *N*. The length of the confidence interval will shorten (regrettably, not as rapidly as we wish, but inversely proportional to $\sqrt{N}$). For example, with $N = 10{,}000$ formula (2.2) provides 0.5 ± 0.01. The connection between frequency and probability of an event can therefore be formulated thus:

Having a sufficiently large number of independent experiments, the frequency of an event will, practically speaking, certainly become as near as desired to its probability.

This statement is called the Jakob Bernoulli theorem or the *simplest form of the law of large numbers*. I have introduced it without proof, but you had hardly doubted it in earnest …

We have thus investigated the meaning of the approach of frequency to probability. One more step is left: to determine approximately the probability of an event given its frequency and estimate the error of that approximation. Formula (2.2), or, if you will, (2.3) will help with the latter task.

Suppose that a large number *N* of experiments was conducted and that the frequency $P^*(A)$ of event *A* is derived. Required is an approximate value of its probability. Denote $P^*(A) = p^*$ and $P(A) = p$ and set approximately that

$$p \approx p^*. \qquad (2.4)$$

Estimate now the practically possible maximal error of this approximate equality by formula (2.2). It will show with confidence level 0.95 how much can the frequency deviate from the probability. But how can we manage it? Formula (2.2) includes the unknown probability *p*, which we indeed wish to estimate.

An absolutely legitimate question! But formula (2.2) only serves for *approximately estimating* the confidence interval. For a *rough* estimation of the error of the probability we may replace the unknown *p* by its approximation, the known frequency *p**. So let us do it! Here is an example.

A series of $N = 400$ experiments provided the frequency of an event $p^* = 0.25$. Choose confidence level 0.95 and determine the maximal practically possible error when assuming that $p = p^*$. Formula (2.2) provides

$$0.25 \pm 2\sqrt{0.25 \cdot 0.75/400} \approx 0.25 \pm 0.043.$$

The maximal practically possible error is 0.043. But if it does not suit us? If that error should not exceed 0.01, say? Increase the number of experiments; but by how much? We will again apply our favourite formula (2.2). Assuming that $p = p^* = 0.25$, we will find the approximate value of the maximal practically possible error and equate it to 0.01:

$$2\sqrt{0.25 \cdot 0.75/\mathrm{N}} = 0.01.$$

Solving this equation we obtain $N = 7500$. And so, for calculating probability with confidence level 0.95 given the frequency of the order of 0.25, and an error not exceeding 0.01, we need 7500 experiments (terrible even to think about it!).

Formula (2.2) or the similar formula (2.3) can also help to solve one more question: is it possible to explain the derived deviation of the frequency from probability by *random causes* or does that deviation indicate that the *probability is not such as we thought*.

Suppose we toss a coin $N = 800$ times and the frequency of heads is 0.52. We suspect that the coin is irregular so that heads appear more often than tails. Is our suspicion warranted? We will start by assuming that everything was in order: the coin is regular, the probability of heads is 0.5 as it should be. Then we will determine the confidence interval at confidence level 0.95 for the frequency of heads. If the determined value 0.52 is within the confidence interval, everything is normal, otherwise we should suspect the coin's regularity.

For the frequency of heads formula (2.2) provides an approximate value 0.5 ± 0.035. The calculated value of the frequency is contained within that interval, so we ought to *exonerate* our coin.

Similar methods are applied for judging whether various deviations from mean values observed in random phenomena are accidental or *significant*. Thus, whether some short measures in a few purchases were random or indicated a systematic deception of customers;

whether the recovery rate of patients heightened accidentally or due to the action of a certain new medicine.

And so, you have learned to determine approximately the probability of events in experiments not reducible to the pattern of chances when issuing from statistical data and even to estimate somehow the ensuing error. But what kind of a science is this theory of probability, will be your possible question. If the probability of an event cannot be determined by formula (1.2), we have to conduct experiments, and more and more of them, until patience and endurance last. Then, we calculate the frequency of the event, equate it to the unknown probability and perhaps estimate the ensuing error. How boring!

You will be completely wrong since such *statistical* derivation of probability is not the only and far from being the main method[20]. Much more important are the indirect rather than direct methods of determining probabilities. They allow us to calculate probabilities of events by issuing from the probabilities of other events, connected with the studied event; to calculate the probability of a compound event by that of a simple, then to go over to simpler events etc. This chain extends to the simplest events after which it cannot be continued further. The probabilities of those simplest events are either calculated by formula (1.2) or derived experimentally, by applying frequencies.

This last-mentioned procedure certainly demands us to conduct experiments or collect data. We should try to use the longest possible chains of events and conduct simplest and cheapest experiments. To attain our goal we therefore ought to get as much as possible information by calculating and as little as possible by experimenting. Indeed, the cheapest components needed for information are paper and the researcher's time.

In the next chapter, we discuss the calculation of probabilities of compound events by those of simple events.

Chapter 3. The Main Principles of the Theory of Probability

I have just emphasised that the main methods of determining the probabilities of events are indirect and consist in issuing from the probabilities of simple events. Here, we deal with these methods. All of them rest on the two proverbial whales, on the two most important principles, *rules* of the theory of probability, those of addition and multiplication of probabilities.

The addition rule. *The probability of the occurrence of whichever of two incompatible events*, *A and B*, *is equal to the sum of their probabilities*

$$P(A \text{ or } B) = P(A) + P(B). \qquad (3.1)$$

Now, is it a theorem or an axiom? Both. It can be rigorously proven for experiments reducible to the pattern of chances. The number of cases favourable for the compound event (A or B) is $m_A + m_B$ etc and the formula (3.1) is therefore often called *addition theorem*.

However, we ought not to forget that for other experiments it is assumed as a *principle*, or an *axiom*. You can convince yourselves in that this theorem/principle is also valid for frequencies.

It can be simply generalized on any number of events. If events A_1, A_2, …, A_n are incompatible, then

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n) = P(A_1) + P(A_2) + \dots + P(A_n). \quad (3.2)$$

There are some important corollaries. First, *if events* A_1, A_2, …, A_n *are mutually incompatible and form a complete group, the sum of their probabilities is unity*. Try to prove this statement yourselves.

Second (a corollary of the corollary): if A is some event and $\bar{A}$ is its contrary event (non-appearance of A), then

$$P(A) + P(\bar{A}) = 1; \quad (3.3)$$

the sum of the probabilities of contrary events is unity. This formula is the foundation of a very common method of *transition to the contrary event*. It often occurs that it is difficult to calculate the probability of some event A, but easy to determine that of $\bar{A}$.

Multiplication rule. *The probability of the combination of two events (of the occurrence of both) is equal to the probability of one of these multiplied by the probability of the other one provided that the first event had occurred*:

$$P(A \text{ and } B) = P(A)\cdot P(B/A). \quad (3.4)$$

Here, $P(B/A)$ is the so-called *conditional probability* of event B calculated under the condition that A had occurred. This formula is a *theorem* as well and can be rigorously proven for the pattern of chances. Otherwise, it is assumed without proof as a *principle* or *axiom*, and it is also valid for frequencies. Note that is absolutely indifferent which event is called the *first*, and which is therefore the *second*. Formula (3.4) can be written as

$$P(A \text{ and } B) = P(B)\cdot P(A/B). \quad (3.5)$$

Ex. 1. An urn contains 3 white and 4 black balls. Two balls are drawn, one after the other. Required is the probability that both are white.

Suppose that both balls are white (events A and B). We ought to find out the probability of their combination. We have formula (3.4) with $P(A) = 3/7$. The second ball is chosen out of the six left, two of them white, so $P(B/A) = 2/6 = 1/3$ and

$$P(A \text{ and } B) = 3/7\cdot 1/3 = 1/7.$$

We have obtained the same result (Ex. 4 in Chapter 1) by a direct enumeration of chances.

But will the solution change if the balls are drawn not successively, but at once? At a glance, it possibly seems that it will indeed change. However, think just a little bit and it will become clear that no change will happen. Indeed. Let us draw the balls at the same time, one with our right hand, the other one, with our left, and call them the first and the second respectively. Will our considerations differ from those applied when solving Ex. 2 [Ex. 5 of Chapter 1]? No, not at all. The probability sought will still be 1/7.

But suppose we draw them both by the same hand? Then, call the ball nearer to the thumb the first one, and stop nagging! But if … Oh, enough is enough, doubting Thomas. You have certainly understood it already.

The multiplication rule becomes especially simple for a special kind of events called *independent*. Two events, A and B, are called independent if the occurrence of one of them does not at all influence the probability of the occurrence of the second one. Or, the conditional probability of event A provided that B had occurred, is absolutely the same as it would be if that restriction was dropped:

$$P(A/B) = P(A). \tag{3.6}$$

Otherwise events A and B are called *dependent*[21].

In our Ex. 1 the events A and B were dependent: the probability of the occurrence of a white ball at the second drawing *depended* on whether the first drawn ball was white or black. But change now the conditions of the problem: return back the first drawn ball, shuffle all of them and draw the second ball. Here, the same events A and B will be independent:

$$P(B/A) = P(B) = 3/7.$$

The notion of dependence/independence of events is very important. Its incomplete understanding often leads to mistakes. Especially beginners tend to forget about dependence of events when it exits, or, conversely, assign some dependence to actually independent events. Let us, for example, ask someone inexperienced in the theory of probability whether heads or tails will be more probable to occur after 10 heads in succession. Almost for sure he will say, Certainly tails. The tosses should sometime compensate each other, tails should arrive sooner or later!

Yes, he will say that and be absolutely wrong. The probability of heads, if only we toss the coin in the usual way, does not at all depend on what happened before. The probability of heads at any toss of a regular coin is 1/2. But of course the appearance of heads 10 times in succession can lead us to suspect that regularity, and we will tend to believe that the appearance of heads is more probable at any toss.

Possibly you will not agree. Well, let us bet. Suppose that a year ago you tossed a coin 10 times and heads appeared invariably. Today, you recalled that curious episode and decided to toss a coin once more. Do you still think that tails will be more probable than heads? You have rather begun to hesitate … but I will now deal the final blow. Let someone else (Pearson, for the sake of definiteness) toss a coin 24 thousand times, and in some 10 consecutive tosses obtain 10 heads. You have recalled that experiment and wish to toss a coin once more. So what is more probable, heads or tails? You have apparently surrendered and admitted that they are equally probable.

Now let us extend our multiplication rule on several events. In the general case of dependent events the probability of one of them is multiplied by the probability of another one provided that the former had occurred, then by the conditional probability of the third one provided that the first two had occurred etc. I will not write out the appropriate formula since this statement is easier to remember when offered verbally.

Ex. 2. An urn contains 5 numbered balls. All of them are drawn one after another. Required is the probability that they will be drawn in their order 1, 2, …, 5.

By the multiplication rule[22]

$$P(1, 2, 3, 4, 5) = 1/5\cdot 1/4\cdot 1/3\cdot 1/2\cdot 1 = 1/120.$$

Ex. 3. Eight separate letters are lying on the table, 2 letters *u*, 3 letters *g* and 3 letters *m*. We pick up three of them one after another. Required is the probability that, being arranged in their appeared order, they form the word *gum*.

By the multiplication rule, $P(\text{gum}) = 3/8\cdot 2/7\cdot 3/6 = 3/56$.

Now, required is the probability that that same word can be formed from the same three letters. Both the conditions of the experiment, and the event itself has changed. We only have to choose letters *g*, *u* and *m* in whichever order. How many cases are there? As many as there are permutations of 3 elements, $P_3 = 3! = 6$. We ought to calculate the probability of each and sum them up. Those probabilities are $2/8\cdot 3/7\cdot 3/6 = 3/56$ and $3/8\cdot 2/7\cdot 3/6 = 3/56$ etc and $3/56\cdot 6 = 9/28$.

The multiplication rule becomes especially simple when the events are independent[23]. We should then multiply not the conditional probabilities, but simply probabilities:

$$P(A_1 \text{ and } A_2 \text{ and } \dots \text{ and } A_n) = P(A_1)\cdot P(A_2)\cdot \ldots \cdot P(A_n). \quad (3.7)$$

The probability of the combination of independent events is equal to the product of their probabilities.

Ex. 4. A shot fires at a target 4 times independently. The probability of a hit is 03. Required is the probability that the 3 first shots fail (–) and the fourth hits the target (+).

For independent events $P(- - - +) = 0.7^3\cdot 0.3 = 0.1029$.

And now a bit more difficult problem.

Ex. 5. Under the conditions of the previous problem required is the probability that exactly 2 shots will be successful.

The demanded aim can be reached in several ways, in as many as there are combinations of 4 elements taken 2 at a time:

$$\binom{4}{2} = 4 \cdot 3/1 \cdot 2 = 6.$$

Here (but not always) the probability of each is the same and equals $0.3^2 \cdot 0.7^2 = 0.0441$ so that the probability sought is $0.0441 \cdot 6 = 0.2646 \approx 0.265$.

According to a general rule useful for solving such problems, the first thing to do is to ask ourselves, in how many incompatible ways can the studied event happen? Then calculate the probability of each and sum them up etc. In the next problem, the probabilities of those different ways are not equal to one another.

Ex. 6. Each of three men shoots once at the same target. Their probabilities of success are $p_1 = 0.4$, $p_2 = 0.5$ and $p_3 = 0.7$. Required is the probability that there will be exactly two hits.

Three ways (combination of 3 elements taken by 1 at a time) lead to success:

$P_1(+ + -) = 0.4 \cdot 0.5 \cdot 0.3 = 0.060$
$P_2(+ - +) = 0.4 \cdot 0.5 \cdot 0.7 = 0.140$
$P_3(- + +) = 0.6 \cdot 0.5 \cdot 0.7 = 0.210$

and their sum is 0.410.

The examples concerning shots and hits about experiments not reducible to the pattern of chances are just as unavoidable and traditional as the classical examples with coins, dice etc about experiments of the other kind. The latter examples do not testify to some special inclination for games of chance just as the former do not indicate some special blood-thirstiness. Their authors just choose the simplest possible illustrations. So endure one more example.

Ex. 7. Required is the probability that, under the conditions of Ex. 4, there will be at least one hit.

There are many ways to achieve the stated aim (event C) and it is certainly possible to calculate the probabilities of each of them and sum them up. But this method is really bad. It is much simpler to transfer from C to the contrary event $\overline{C}$ which occurs only in one way:

$$P(\overline{C}) = 0.7^4 \approx 0.240, \text{ so that } P(C) = 0.760.$$

If the contrary event has a lesser number of ways to occur than the event sought, transfer to it. One of the almost sure indication for such a transfer is the presence of the words *at least* in the formulation of the problem.

Ex. 8. A group of n people unknown to each other is formed. Required is the probability that at least two of them have the same date (day and month) of birth (event C).

We assume that birthdays fall on each day of the year with the same probability[24]. Now, *at least* puts us on guard: will not it be better to transfer to the contrary event? Indeed, C has so many ways to occur, that even to think about them makes us feel creepy all over. On the other hand, $\overline{C}$ is much more modest and its probability can be obtained very simply.

Call one participant of the group the *first* person. He can be born on any day of the year (probability 1). Now choose arbitrarily a second person who can be born on any day except that on which the first one was born (probability 364/365) etc[25]. Therefore,

$$P(\bar{C}) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365-(n-1)}{365}\ ,\ P(C) = 1 - P(\bar{C}). \qquad (3.8)$$

A curious feature of this problem consists in that with a (even rather modestly) increasing n the event C rapidly becomes almost certain. Thus, for $C = 50$ formula (3.8) provides $P(\bar{C}) \approx 0.03$, $P(C) \approx 0.97$. With a high level of confidence (0.97) event C can be considered certain!

This uncomplicated calculation can help you, if you so desire, to become a *magician*. Maintain that in a gathering of 50 or of a bit larger number of people[26] whose birthdays are unknown to you there are those whose birthdays coincide. Take a sheet of paper number 31 rows (1, 2, …, 31) and 12 columns (January, February, …, December), ask each person to mark his birthday in the appropriate cell, and let me see when two marks coincide. But suppose no such coincidence takes place? Oh, although being sick at heart, confidently tell them that it will happen for sure. Actually, you know that your prediction will not come true absolutely certainly: with a very low probability it will fail. Well, you are risking.

And now we turn to serious matters by solving an important general problem often occurring in most various forms.

Problem 1. Event A occurs with probability p in each of n independent experiments. Required is the probability that it occurs at least once (event C).

The magic words *at least* send us to the contrary event which is simpler and can only occur in one way. Then, by the multiplication rule for independent events,

$$P(\bar{C}) = (1 - p)^n,\ P(C) = 1 - (1 - p)^n. \qquad (3.9a, 3.9b)$$

Pay attention to formula (3.11b): it is being applied for solving many practically important problems.

Ex. 9. The probability of detecting an artificial space object by a single radiolocation sweep is $p = 0.1$. Required is the probability that such an object will be detected after 10 independent sweeps.

By formula (3.9b)

$$P(C) = 1 - (1 - 0.1)^{10} = 1 - 0.348 = 0.652.$$

Ex. 10. A technical device consists of 7 elements each failing independently from the others with probability 0.05[27]. A failure of even one element leads to some breakdown. Required is the probability of such an event (C).

By formula (3.9b)

$P(C) = 1 - (1 - 0.05)^7 = 0.305.$

Surprise! That probability is higher than 30%. The reliability (the probability of being in good repair) of each element should be urgently heightened! Note that the failure of each had a rather low probability 0.05. When considering that probability thoughtlessly it is possible to disregard it and declare that the failure of the device was practically impossible. We have done the same when predicting the results of coin tossing but here we have something altogether different. First, there are 7 elements rather than 1, and the failure of *at least one* is not unlikely at all. In addition, the consequences of thoughtlessness are not in the least harmless.

It is interesting to note that stochastic calculations sometimes lead to unexpected results as though contradicting common sense. Here is an (apparently) amusing example.

Ex. 12. Two hunters, Simon and Georgy, saw a bear and shot at him at the same time. They killed the bear and found only one hit. It is more likely that it was Simon who killed him since he was an old hand at hunting and considering the shot's distance would have hit the bear with probability $p_1 = 0.8$. Georgy, however, is a young and less experienced hunter, and for him the corresponding probability was only $p_2 = 0.4$. The hunters sold the bear's fell and required is how they should share the earned money.

You will probably wish to share that money in the proportion to these probabilities, to give Simon 2/3 of the money, and 1/3 to Georgy. Just imagine, however, that you are wrong! To convince you, I change the conditions of the problem and let $p_1 = 1$ and $p_2 = 0.5$. Will the fell belong to Simon? Certainly, since he could not have missed. You, however, would have shared the money just as previously, in the same proportion. So something is wrong, but what exactly?

You have not taken into account that one of the hunters failed, so let us now properly solve this problem. The hit could have happened in two ways:

A_1: Simon hit the bear, but Georgy did not.
A_2: Georgy hit the bear, but Simon did not.

By the multiplication rule

$P(A_1) = 0.8 \cdot 0.6 = 0.48$, $P(A_2) = 0.4 \cdot 0.2 = 0.08$.

And so, the earned money should be shared in proportion to these probabilities, 0.48 and 0.08. But then, Simon, who got the lion's share will likely spend it on treating them both to a hunter's meal at a campfire.

Chapter 4. Random Variables

In this chapter you will become acquainted with a new and very important notion of *random variable*. In Chapter 1 you learned the simplest method of calculating probabilities of events, by directly computing the fraction of favourable cases. Immediately after that, however, you became disappointed: it occurred that that method is applicable far from always, only in those comparatively rare problems in which the experiment is reducible to the pattern of chances, i. e., when its possible outcomes are symmetrical.

Nevertheless, in the next chapter you learned to derive approximately the probabilities of events by their frequencies without any restrictions imposed on the experiment. And again, immediately afterwards you were disappointed once more: it occurred that such derivations did not constitute the main stochastic method. Finally, in Chapter 3, according to its title, you came across such main methods. Here we are, you believe, I have now succeeded to reach the very foundations, the very essence. No more disappointments!

Alas, one more disappointment (the last one) is still lying in wait. The point is that according to the contemporary theory of probability, an event with which we had to deal until now is not its main notion. And what is the main notion, you will ask me in a burst of indignation, and what the deuce was I doing, compelled to get acquainted with inferior armoury? Unfortunately, I meekly answer, without that armoury it is impossible even to approach the modern arsenal of random variables (RV).

This chapter is indeed devoted to that notion, to the main notion of the contemporary theory of probability, to RVs, to their varieties and methods of describing and dealing with them. As compared with events, we will discuss them in lesser detail, consider them rather in a descriptive manner since the mathematical apparatus, had we applied it, would have been more complicated possibly repelling the beginner. Indeed, it is exactly our aim to simplify his *first steps*. And so,

a RV is a variable, which, as a result of an experiment, can take one or another value, unknown beforehand.

As always in the theory of probability, this statement is somewhat obscure; it includes some indefiniteness and unknowns[28]. For mastering the definition, just get acquainted with it, so let us consider some examples.

1. We toss two coins. The number of appearing heads is a RV with possible values 0, 1, 2, and we do not know which of them will be realized.

2. A student is examined. His mark is a RV with possible values 2, 3, 4, 5[29].

3. There are 28 students in a group. The number of those failing to appear on a certain date is a RV with possible values 0, 1; …, 28. Impossible! All of them cannot be taken ill (or play truant) at once. Yes, such an event is practically impossible, but who told you that all the values of a RV ought to be equally probable?

All the RVs in those examples belong to the so-called *discrete* type which means that their possible values are separated by some intervals. When shown on the number axis, they are represented by *separate points*. There also exist *continuous* RVs whose values completely fill some interval on that axis. The boundaries of such intervals are either definite and clear or vague. Here are a few examples.

4. The time interval between two consecutive failures of a computer. The values of this RV completely fill some part of the number axis.

The left boundary (0) of that part is quite clear, but the right one is indefinite and can only be established by experiments.

5. The weight of a freight train.

6. The water level at flood-time.

7. The error of weighing something by an analytical balance. Unlike the RVs above, this RV can take both positive and negative values.

8. The specific gravity of milk in its sample.

9. The time daily spent by a school student of a certain form on watching television.

I emphasize that, when discussing a RV, it is necessary to specify *the essence of the corresponding experiment*. Thus, in the fourth example we should specify the type of the computer, its age and the conditions of its work. In the seventh example, it is necessary to specify the balance and the set of weights.

We will not invariably mention this condition. Note that all the continuous RVs [all their values] can only be measured in some units (minutes, centimetres, tons) and in the strict sense they are discrete. For example, it is meaningless to measure a person's stature more precisely than to within 1 *cm*. And so, stature is in essence a discrete RV with values separated by intervals of 1 *cm*. The number of such [possible] values is very large and they are situated very closely[30]. It is then more convenient to consider that the RV belongs to the continuous type. We will denote RVs by capital letters, and their possible values, by the same small letters.

Let us now have a RV taking some values. They naturally are not equally probable; some are more, and some are less probable. The *law of distribution* of a RV is any function describing the distribution of probabilities of its values. I will only acquaint you with some of the simplest laws. The law of distribution of a discrete RV can be written down as a table with two rows of its possible values $x_1, x_2, \ldots, x_n$ and their probabilities $p_1, p_2, \ldots, p_n$. Each p_i is simply the probability that the RV takes the value x_i and their sum is apparently unity:

$$p_1 + p_2 + \ldots + p_n = 1.$$

This unity is somehow distributed among the values of the RV, hence the term *distribution*.

Ex. 1. Three independent shots at a target are made and the probability of each hit is 0.4. The discrete RV is the number of hits. Show its law of distribution (its possible values with their probabilities).

We have

$p_0 = P(-\ -\ -) = 0.6^3 = 0.216$

$p_1 = P[(+\ -\ -) \text{ or } (-\ -\ +) \text{ or } (-\ +\ -)] = 3\cdot0.6^2\cdot0.4 = 0.432$

$p_2 = P[(+\ +\ -) \text{ or } (+\ -\ +) \text{ or } (-\ +\ +)] = 3\cdot0.4^2\cdot0.6 = 0.288$

$p_3 = P(+\ +\ +) = 0.4^3 = 0.064$

The sum of these probabilities is indeed unity.

Ex. 2. A sportsman is attempting to throw a ball into the basket. At each independent attempt the probability of his success is p. The discrete RV is here the number of his attempts continuing until success.

The possible values of the RV are here $x_1 = 1, x_2 = 2, \dots, x_k = k$ (theoretically, k can be infinite). Now, $p_1 = p$, but for determining p_2 we have to consider the combination of two events, of the failure at the first attempt and success at the second one, $p_2 = (1 - p)p$. Similarly, $p_3 = (1 - p)^2p$ and in general $p_i = (1 - p)^{i-1} p$. These probabilities p_i form a geometric progression with ratio $1 - p$ so that the corresponding distribution is called *geometrical*.

Let us see now how to characterize the distribution of probabilities for a continuous RV. The table with two rows (see above) cannot be formed; it is impossible even to fill its upper row, that is, to list all the possible values of the RV one after another: between any two of them other values will inevitably exist[31]. Another difficulty consists in that any separate value of a continuous RV has probability 0. Yes, exactly so, I am not mistaken and will try to convince you.

Suppose you are on a pebbly beach and are interested in the RV, in the weight of an isolated pebble. So let us weigh the pebbles. We begin with a reasonable precision, 1 *g*, and consider the weight of each pebble equal to 30 *g* (say) if it is 30 *g* to within 1 *g*. We will obtain the frequency of that weight although we neither know, nor need to know it.

Let us then weigh to within 0.1 *g*, so that some pebbles supposed to weigh 30 *g* will now be left out. The frequency of the event X [x rather than X] = 30 *g* will decrease. By how much? It will become about 10 times less. And now we will weigh to within 1 milligram and the frequency will become 100 times less. But frequency is the own sister of probability and approaches it as the number of experiments (of the sufficiently numerous pebbles on the beach) increases. So what value should we assign to the probability that the pebble's weight is exactly 30 *g*, not a bit more, not a bit less? Obviously, zero; what can we do otherwise?

You are perplexed, perhaps indignant since you certainly remember that zero probabilities mean *impossible events* whereas a continuous RV can take some value x, so how can its probability be zero? Let us remember everything well and truly. Yes, we did state that *the probability of an impossible event* is zero. But had we ever maintained that any event with zero probability is impossible? No, not at all. And now we had to acquaint ourselves with possible events having zero probabilities.

Don't hurry, let us reflect awhile. Forget the theory of probability for a moment and imagine some plane figure with area S. Choose any point inside; what is its area? Apparently, zero, but the figure obviously consists of points each having a zero area whereas $S > 0$. This paradox is not surprising since you got used to it, and now you ought to become accustomed to the fact that, having a continuous RV, the probability of choosing each isolated point is exactly zero[32].

So how then can we discuss a *distribution of probabilities* for such RVs if each of its values has the same probability, zero? You are quite right. It is senseless to consider the distribution of probabilities among *separate values* of a continuous RV. Nevertheless, a distribution does exist. For example, no one doubts that 170 *cm* is more probable than 210 *cm* for a man's stature although both are possible.

We ought to introduce now a new important notion, *density of probability*. Density of a substance is sufficiently known in physics, it is the weight of its unit volume. But if the substance is heterogeneous? Then we can only consider its *local* density. The same happens in the theory of probability. We consider the local density (the probability of the RV's unit length at point x).

The density of probability of a continuous RV is the limit of the ratio of the probability of its occurring in a small interval adjacent to point x to the length of that interval if that latter tends to disappear.

Density of probability is easily derived from its similar notion of *density of frequency*. Consider a continuous RV (someone's stature or the weight of a pebble). First of all, conduct a series of experiments with that RV which will take some value in each of them. Thus, measure the stature of each person from a group of people or weigh many pebbles. We are interested in the distribution of the probability of our RV. Separate the range of its values into some intervals; for example, let them be 150 – 155, 155 – 160, …, 195 – 200 *cm*. Count the number of the values in each interval[33], divide these numbers by the total number of experiments and by the lengths of the intervals (which do not have to be the same), and thus obtain the *density of frequency*.

When having at our disposal enough statistical data (of the order of a few hundred, or more, which is better) we ensure a sufficiently good notion about the distribution of the RV, about its density. It is expedient to begin here by constructing a special bar chart consisting of rectangles whose areas are equal to the frequencies of the separate intervals (and whose heights are therefore the densities of the frequencies). The area of a bar chart is obviously unity. With an increasing number of experiments the intervals of the RV's range of values can be shortened, the *steps* of the chart will then smoothen and the chart will approach some fluent curve, the *curve of distribution*. Ordinates will become the densities of probability rather than of frequency, and the complete area restricted by that curve just like the area of its sister, of the bar chart, will be unity.

The probability of the RV being in some interval (a, b) will then be equal to the area of the figure resting on that interval. Denote the density of probability by $f(x)$, then that probability will be represented by integral

$$P(a,b) = \int_a^b f(x)dx. \qquad (4.1)$$

Sparing neither time nor money we can determine $f(x)$ by experimental data as precisely as desired. But is the game worth the candle? Do we need to know that function *absolutely precisely*? Very often we do not and are rather satisfied by an approximate notion of the RV's law of distribution. Indeed, all stochastic calculations are in essence approximate so that the number of experiments can be rather modest, 300 – 400, say (!), and sometimes less.

It is possible to construct a bar chart and then smooth it by some fluent curve (but certainly keeping the area *under* the curve equal to unity). The theory of probability has at its disposal an entire set of such curves. Some of them are in a sense better than others; for example, when choosing properly, the integral (4.1) can be obtained more easily, or calculated by appropriate tables. The manner in which the RV had originated can prompt us to choose a certain type of distributions according to theoretical considerations. I will not dwell on such details since this is a special topic. However, I emphasize that *indirect* methods of determining laws of distribution are more important here than the *direct* approach. They allow us to derive the distribution of a RV not directly, by experiments, but by issuing from other RVs somehow connected with it.

The so-called *numerical characteristics* of RVs play a large part in realizing these indirect methods. These are numbers describing some of their distinguishing properties. For example, the mean value in whose vicinity occur the random deviations [of the values of a RV]; the magnitude of those deviations (as though the degree of randomness of the RV) and some other indications. Many problems can be solved by applying those characteristics without or almost without resorting to the laws of distribution. I am only acquainting you with two (but the most important two) numerical characteristics, the *expectation* and the *variance*.

Expectation EX of a discrete RV X is the sum of the products of all its possible values by these probabilities:

$$\mathrm{E}X = x_1p_1 + x_2p_2 + \ldots + x_np_n = \sum_{i=1}^{n} x_i p_i. \qquad (4.2, 4.3)$$

Formulas (4.2) and (4.3) show that EX is the generalized, weighted mean of all possible values of X with weights being the corresponding probabilities. If the RV has infinitely many possible values, the sums (4.2) and (4.3) will include infinitely many products.

The expectation or mean value of a RV is as though its *representative* which can replace it for rough estimations. Actually, we always do it when randomness is not taken into account.

Ex. 3. Determine the expectation of the RV in Ex. 1 (the number of hits after 3 shots).

By formula (4.2)

$$\mathrm{E}X = 0\cdot 0.216 + 1\cdot 0.432 + 2\cdot 0.288 + 3\cdot 0.064 = 1.2.$$

Expectation is also introduced for continuous RVs but then the sum in formula (4.3) is naturally replaced by an integral:

$$\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)dx. \qquad (4.4)$$

Here, $f(x)$ is the density of the RV's probability.

Let us now briefly discuss the expectation, its meaning and *genealogy*. Probability has its own sister, frequency, and expectation has its own brother (sister? relative?), the arithmetic mean of the observational results. Just as frequency approaches probability as the number of experiments increases, that mean of the observational results of a RV approaches expectation.

Let us prove it for discrete RVs**34**. I think that you will gladly accept it on trust for continuous RVs. Suppose that N experiments were made and value x_1 occurred M_1 times, x_2, M_2 times, etc. The arithmetic mean $\overline{X}$ of the values of X is

$$\overline{X} = \frac{x_1M_1 + x_2M_2 + ... + x_nM_n}{N} = \sum \frac{x_iM_i}{N}.$$

But $M_i/N = p^*_i$ is the frequency of the event $X = x_i$; with practical certainty it approaches probability p_i as N increases. Therefore,

The arithmetic mean of the observed values of a RV will with practical certainty approach without bound its expectation as the number of the experiments increases.

This statement represents one of the forms of the *law of large numbers*, a theorem very important for practically applying the theory of probability. In a long series of experiments, the unknown probability of an event can be approximately determined by its frequency, and just the same the expectation of a RV can be approximately assumed as the arithmetic mean of its observed values

$$\text{E}X \approx \overline{X}. \qquad (4.5)$$

I especially note that for such calculations we do not at all need to know the law of distribution of the RV. We just calculate the mean of all of its observational results. One more remark. For determining with a sufficient precision the expectation of a RV we do not at all need the same number of experiments (of the order of a few hundred) as for constructing a bar chart. Several dozen are sufficient.

Introduce now the second most important numerical characteristic of a RV: its *variance*, var X, which describes the scattering of its values around the mean. The larger the variance, the *more random* is the RV. Here is how it is calculated. The mean (the expectation) is subtracted from each possible value of the RV, the differences are squared, multiplied by the corresponding probabilities and finally all such products are summed up:

$$\text{var}\, X = \sum_{i=1}^{n} (x_i - \text{E}x_i)^2 p_i. \qquad (4.6)$$

But why square the differences? To get rid of the signs (plus or minus). Instead, we could have certainly introduced absolute values of those differences but the resulting measure will then be less convenient for calculating and dealing with.

Ex. 4. Determine the variance of the number of hits X in Ex. 1.

By formula (4.6)

var $X = (0 - 1.2)^2 \cdot 0.216 + (1 - 1.2)^2 \cdot 0.432 +$
$(2 - 1.2)^2 \cdot 0.288 + (3 - 1.2)^2 \cdot 0.064 = 0.72.$

Formula (4.6) is not however the best for calculating. It is usually convenient to apply formula

$$\text{var}\,X = \mathrm{E}[X^2] - (\mathrm{E}X)^2. \tag{4.7}$$

The variance of a RV equals the expectation of its square less the square of its expectation.

Formula (4.7) is easily derived by identical transformations, but we will rather confirm its validity by calculating the previous example anew:

$$\text{var}\,X = 0^2 \cdot 0.216 + 1^2 \cdot 0.432 + 2^2 \cdot 0.288 + 3^2 \cdot 0.064 - (1.2)^2 = 0.72.$$

For continuous RVs variance is calculated similarly, by formula (4.6), but with the sum being naturally enough replaced by an integral:

$$\text{var}\,X = \int_{-\infty}^{\infty} (x - \mathrm{E}x)^2 f(x) dx. \tag{4.8}$$

It is usually more convenient, however, to apply formula (4.7), again with a similar replacement:

$$\text{var}\,X = \int_{-\infty}^{\infty} x^2 f(x) - (\mathrm{E}x)^2 dx. \tag{4.9}$$

For an approximate derivation of the expectation we did not need to know the appropriate law of distribution, and now we can approximately directly calculate the variance by observations, by the deviations of the observed values of the RV from their mean:

$$\text{var}\,X \approx \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{X})^2. \tag{4.10}$$

Here, k is the number of the experiment, x_k, the corresponding value of the RV and N is the number of experiments. Again, it is more convenient to apply a formula similar to (4.7)

$$\text{var}\,X \approx \frac{1}{N}\sum_{k=1}^{N} x_k^2 - \bar{X}^2. \tag{4.11}$$

Formulas (4.10) and (4.11) can be applied for calculating a rough estimate of the variance without having very many experiments (better something than nothing). In mathematical statistics, a *correction for a small number of experiments* is usually made by multiplying the result obtained by $N/(N - 1)$. We certainly may do so, but this correction is not very important since with a small number of experiments nothing good will result by treating them anyhow. And with a large N the correction is close to 1.

Variance as a measure of scattering has an unpleasant feature: its dimensionality, see formula (4.6), is equal to the square of X's dimensionality. For example, if X is expressed, say, in minutes, its variance is in *square minutes*, which is not very clear. For avoiding this circumstance a square root is extracted of the variance and a new measure of scattering thus emerges, the so-called mean square (or standard) deviation

$$\sigma_x = \sqrt{\text{var}\, X}. \qquad (4.12)$$

This is a very transparent and convenient measure of scattering. It immediately provides a notion about the range of the oscillations of a RV from its mean. For RVs mostly occurring in practice it can be stated *with practical certainty that they do not deviate from their expectations more than by* $3\sigma_x$.

Confidence level depends on the RV's law of distribution, but it is rather high in each problem, if not otherwise artificially designed. The statement above is called the three-sigma rule. Therefore, when being able to determine somehow both numerical characteristic of a RV, its expectation and variance, we immediately get an approximate idea about the range of its possible values.

You may ask here: we determined those parameters experimentally, so why cannot we find that range the same way? Yes, you are quire right if these characteristics are indeed found directly by experimenting. But that (direct) approach does not constitute the main method of determining the numerical characteristics. We say it once more: the main methods are indirect, those that allow us to determine the numerical characteristics of RVs by characteristics of other RVs connected with the former.

In such cases we apply the *main rules* of dealing with these parameters; we will formulate now (certainly without proof) some of those rules.

1. The expectation of a sum of RVs is equal to the sum of the expectations of the summands.

2. The variance of a sum of independent RVs is equal to the sum of the variances of the summands.

3. A non-random factor c can be taken out from the sign of expectation:

$$\text{E}(cX) = c\text{E}X.$$

4. When squared, a non-random factor c can be taken out from the sign of variance

$$\text{var}(cX) = c^2\text{var}X.$$

These rules although perhaps not the last one seem natural. I will now show the validity of that last one by the following example. Suppose you double the RV X. Its expectation obviously doubles as well and the same happens to the deviation of a separate value of the RV from its mean and the square of that deviation becomes quadrupled.

The small number of rules provided above is nevertheless sufficient for solving some interesting problems. Indeed:

Problem 1. N independent experiments are made. Event A occurs in each with probability p. Required is the expectation and variance of the random number X of experiments in which A appears.

Represent X as a sum of N RVs:

$$X = X_1 + X_2 + \ldots + X_N. \qquad (4.13)$$

Here, $X_k = 1$ if A occurred in the k-th experiment and $X_k = 0$ otherwise. By rule No. 1

$$\mathrm{E}X = \mathrm{E}X_1 + \mathrm{E}X_2 + \ldots + \mathrm{E}X_N. \qquad (4.14)$$

The experiments are independent and so therefore are the X_k. By rule No. 3

$$\mathrm{var}X = \mathrm{var}X_1 + \mathrm{var}X_2 + \ldots + \mathrm{var}X_N. \qquad (4.15)$$

Now we have to determine the expectation and variance of each X_k. They are discrete RVs with two possible values, 0 and 1 having probabilities $(1 - p)$ and p respectively. The expectation of each is

$$\mathrm{E}X_k = 0\cdot(1 - p) + 1\cdot p = p.$$

By formula (4.7) their variances are

$$\mathrm{var}X_k = 0^2\cdot(1 - p) + 1^2\cdot p - (\mathrm{E}X)^2 = p - p^2 = p(1 - p).$$

Apply now formulas (4.14) and (4.15):

$$\mathrm{E}X = Np, \ \mathrm{var}X = Np(1 - p).$$

Problem 2. Under the same conditions approximately determine the range of the practically possible values of the RV P^*, of the frequency of A.

By definition, frequency is the number X of the occurrences of the event divided by N. By rules 3 and 4 we have

$$\mathrm{E}P^* = \mathrm{E}(X/N) = (1/N)\mathrm{E}X = Np/N = p.$$
$$\mathrm{var}P^* = \mathrm{var}\,(X/N) = (1/N^2)\mathrm{var}X = Np(1 - p)/N^2 = p(1 - p)/N.$$
$$\sigma_{P^*} = \sqrt{p(1-p)/N}.$$

Apply now the three-sigma rule to determine approximately the range of the practically possible values of P^*:

$$p \pm 3\sigma_{P^*}.$$

Well, isn't it our old friend, you will exclaim if you have attentively read the previous material, This very formula appeared for the confidence interval with confidence level 0.997 and estimated the value of the frequency of an event given a large number of experiments. And, along with it (and even considered preferable) another formula with coefficient 2 rather than 3 was recommended, the formula which is realized with probability 0.95. Yes, but wherefrom did the probabilities 0.997 and 0.95 arrive?

Just a minute, be patient. You should become acquainted with a very important law of probability, the so-called *normal law*. Consider a continuous RV X. It is normally distributed if the density of its probability is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-m)^2}{2\sigma^2}]. \tag{4.16}$$

[…] This law depends on two parameters, m and σ. The first, as you probably surmised, is the expectation of X, and the second is its mean square deviation. Change m and the curve, without changing its form, will shift in one or another direction along the x-axis. Change σ and the curve will change its form: increase σ and it spreads out; decrease it, and the curve becomes needle-shaped.

The special role played by the normal law is connected with one of its remarkable properties. When summing up a large number of independent (or weekly dependent) RVs, comparable with respect to the order of their variances, *the law of distribution of the sum will be close to the normal law* (the closer, the more is that number) *whichever are the laws of the summands*.

This is a rough formulation of the very important so-called *central limit theorem*. It is known in many various forms differing in the underlying conditions which the initial RVs ought to satisfy. In practice, very many RVs are formed *by summing up* and are therefore distributed normally or almost so. For example, the errors of all types of measurements are sums of many *elementary* and practically independent errors, the effects of their own causes[35].

As a rule, the errors of firing obey the same rule as do the deviations of the voltage in an electrical grid from its nominal value, total payments made by an insurance office during a long period, the total time of a computer being out of service during a year etc.

Such an interesting RV as *the frequency of an event A* in a large number of experiments also has a law of distribution close to the normal. Indeed,

$$P^* = (X_1 + X_2 + \dots + X_N)/N$$

where X_k is a RV with values 1 if *A* occurred in the *k*-th experiment and 0 otherwise. The proof is obvious since *P** is the sum of a large number of independent terms having the same variance

$$\text{var}(X_k/N) = (1/N^2)p(1-p).$$

Since the normal law often occurs in practice, the probability of a RV thus distributed being within (*a*, *b*) is calculated time and time again. Now, the integral (4.16) cannot be expressed by elementary functions and has to be calculated by tables of the function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x \exp(-t^2/2)dt.$$

Here is a fragment from such a table […]. For $x \geq 4$ we may assume that, to within the fourth digit after the decimal point, $\Phi(x) = 0.5000$. It should also be bourn in mind that $\Phi(x)$ is an odd function, $\Phi(-x) = -\Phi(x)$.

The probability of the event mentioned above is

$$P(a, b) = \Phi[(b-m)/\sigma] - \Phi[(a-m)/\sigma]. \quad (4.17)$$

Ex. 5. Determine the probability that RV *X* having a normal distribution with parameters *m* and σ will deviate from its expectation not more than by 2σ; by 3σ.

By formula (4.17) and the table [omitted from the translation]

$$P(m-2\sigma,\ m+2\sigma) = \Phi[\frac{m+2\sigma-m}{\sigma}] - \Phi[\frac{m-2\sigma-m}{\sigma}] =$$
$$\Phi(2) - \Phi(-2) = \Phi(2) + \Phi(2) = [\ldots] \approx 0.95.$$

For the second case [the probability is approximately 0.997.]

So here finally they are, those confidence probabilities for the frequency of an event, which, along with the corresponding confidence intervals, had been discussed in Chapter 2. We had to go a long way!

Ex. 6. A freight train has $N = 100$ cars. The weight of each is a RV with expectation $m_q = 65$ ton and mean square deviation $\sigma_q = 9$ ton. A locomotive can pull a train weighing up to 6600 ton, otherwise a second locomotive is needed. Required is the probability that one locomotive is sufficient.

The weight *X* of the train is the sum of 100 RVs Q_k with the same expectations and variances $\sigma_q^2 = 81$. By summing up the expectations and variances we get

$$EX = 100\cdot 65 = 6500,\ \text{var}X = 100\cdot 81 = 8100,\ \sigma_X = 90.$$

We demand that *X* does not exceed 6600. We may assume that it is normally distributed, and by formula (4.17)

$$P(0, 6600) = \Phi[(6600-6500)/90] - \Phi[(0-6500)/90] =$$

[…] ≈ 0.887.

And so, one locomotive is sufficient with probability ≈ 0.887. Suppose now that $N = 98$. Calculate yourselves and the result will likely surprise you: the probability sought is 0.99, that is, practically certain. Only two cars have been uncoupled!

You see now how curious can the problems be when a large number of RVs have to be summed up. Here, however, a question naturally arises: how much is many? How many RVs should we sum up for the law of distribution of the sum to become *normalized*? It depends on the laws of distribution of the summands which should be studied at least in the first approximation. There exist such intricate laws that very, very many summands are needed. To repeat: whatever will those mathematicians devise! However, nature does not intentionally play mean tricks. A sufficient number of summands for the normal law to become applicable, especially if they have the same distributions, is usually 5 – 6, or, well, 10; well, really, 20.

The rapidity with which the law of distribution of a sum of the summands having the same distribution is being *normalized* can be illustrated by an example. You will again have to believe me on trust; I did not yet deceive you. Suppose we have a continuous RV with a constant density on interval (0, 1). The curve of distribution becomes a segment, so unlike the normal law! Sum up two such (independent) RVs, and the density is now represented by triangular law. It does not resemble the normal distribution, but we are moving in the right direction, For the sum of three such uniformly distributed (!) RVs the curve of distribution consists of three parabolic segments and *awfully* resembles the normal law. And when summing up 6 uniformly distributed RVs no one will be able to say that the resulting curve is not normal.

This is the foundation of the commonly applied method of obtaining a normally distributed RV. When simulating random phenomena by a computer it is sufficient to sum up 6 such RVs existing on interval (0, 1). Nevertheless, we should not be excessively carried away and declare at once that the normal law is the distribution of the sum of several RVs but rather somewhat cautiously resort to this rule. At least in the first approximation we ought to study their distributions. If, for example, they are very asymmetric, or if the probability of an event occurring in each experiment is very high or very low, a large number of summands can be necessary.

Incidentally, here is a practical rule allowing us to find out whether the normal law may be assumed for a frequency. Construct as shown above the confidence interval for it with confidence level 0.997

$$p \pm 3\sqrt{p(1-p)/N}.$$

If it is entirely (with both ends!) situated within some reasonable boundaries[36] we may assume a normal law, but not if one of the boundaries is beyond interval (0, 1). For approximately solving our problems in latter cases we may apply the so-called Poisson distribution. However, the pertinent subject as well as a study of other distributions (a great many of them are applied in the theory of probability) is not here possible.

As a result of reading my unpretentious booklet you have apparently gained some understanding about the essence of that theory and its scope. You possibly came to loath it resolutely and then your first acquaintance with it will also be the last one as well. Too bad, but it cannot be helped. There are people (and even mathematicians) who inherently cannot bear the theory of probability. But it is also possible that the subject, the methods and the possibilities of the theory did interest you (which was the aim of my booklet). Then get deeper acquainted with it. I will not conceal from you that that will demand much more mental efforts than the *first steps* did. It will be more difficult but more interesting as well. Not without reason we say that the roots of study are bitter but its fruits are sweet. I am wishing you the sweetest fruits!

Notes

1. The author time and time again applies the term *phenomenon* instead of *event* and prefers *experiment* to *trial*. She calculates with an excessive number of digits after the decimal point.

2. This definition is unfortunate: probabilities are not mentioned although they inevitably appear. The same remark applies to the definitions of a random variable in Chapter 4.

3. There had indeed been such a lottery in the USSR.

4. Here and below the term *likelihood* is applied as stated: *or probability*.

5. De Morgan (1864) believed in negative probabilities and in those higher than unity. Worse (Sophia De Morgan 1882, p. 147), in a letter of 1842 he stated that the tangent and cotangent of infinity are equal to $\pm\sqrt{-1}$.

6. Those few readers who hesitate before agreeing with that answer are faultfinders. Sometimes such behaviour testifies to deep thinking, but more often indicates bad temper. E. V.

Homogeneity of the coin is also necessary.

7. Otherwise (theoretical) probability does not exist.

8. No example is provided and the statement is unfounded.

9. The author did not say that she only discussed subjective probabilities so that her conclusions are barely useful. Poisson (1837, § 11) proved that the subjective probability of drawing a white ball from an urn containing white and black balls in an unknown proportion was 1/2. According to information theory such a probability is tantamount to complete ignorance. The same conclusion applies to the celebrated Bertrand problem about a random chord: after more than a century of discussions commentators largely agreed that the probability of its being shorter than a side of an equilateral triangle inscribed in the circle is 1/2!

10. Just like in Note 8, no example is provided. The statement is unfounded and furthermore doubtful.

11. In the sequel, I drop the adjective *relative*.

12. The author indirectly introduces degrees of randomness (also in Chapter 4). See a similar approach in Chaitin (1975) who had nevertheless connected them with complexity rather than range.

13. The experiment is due to Weldon who rolled 12 dice 26 306 times. Pearson (1900) only discussed it as did Markov (1924, pp. 349 – 353).

14. See Note 12.

15. Under some special conditions (for example, during and after wars) this frequency can deviate from a stable long-term mean. The causes of these deviations are not yet ascertained. E. V.

16. Governing mortality?

17. Even now the existence of organic life on Mars is doubtful.

18. Games of chance are very important at least methodically.

19. The celebrated scientist D'Alembert is said to have given lessons in mathematics to a very slow-witted and very noble pupil who was unable to understand a certain proof. Becoming desperate, D'Alembert cried out: *Upon my word, this theorem is true*. The pupil answered: *But why did not you tell me that from the beginning? You are a nobleman, and I am a nobleman. Your word is quite sufficient for me*! E. V.

20. Statistical probability remains extremely important in statistics itself. Moreover, when solving any practical problem, the researcher has to issue from statistical data.

21. It is easy to prove that dependence and independence are always mutual, i. e., that $P(A/B) = P(A)$ involves $P(B/A) = P(B)$. Show it yourselves by applying the two forms of the multiplication rule, (3.4) and (3.5). E. V.

22. Either here or below I do not introduce special notation. The less symbols are applied, the better it is. E. V.

23. Several events are called independent if none of them depends on any combination of the other ones. For independence of events in their totality independence of their pairs is not sufficient. Intricate examples can be devised of such events which prove the above statement. Trust mathematicians to invent whatever they wish! E. V. See also Rumshitsky, § 1.4 and Note 5.

24. It is not exactly true, but that assumption may be introduced as a first approximation. E. V.

25. This formula is only valid at $n \le 365$; for $n > 365$ we obviously have $P(\bar{C}) = 0.$ For the sake of simplicity we disregard leap years (and birthdays occurring on February 29). E. V.

See an approximate calculation of probability $P(C)$ in Feller (1950/1964, § 2.3).

26. When n is much larger than 365, the experiment becomes barely effective. E. V.

27. That probability obviously depends on the time interval involved.

28. It is really unfortunate to maintain that statements pertaining to the theory of probability are always obscure.

29. According to a long-standing Russian tradition the highest mark (5) corresponds to the largest number, and the lowest mark (2), to the least number.

30. Why is the number of possible discrete measurements of stature *very large*? The same statement is repeated below.

31. For those acquainted with the set theory it can be added that the number of these values is uncountable. E. V.

32. There also exists a special, so-called mixed type of RVs: in addition to a dense interval of possible values having zero probabilities they have separate, special values with positive probabilities. I do not consider such RVs, but their existence ought to be known. E. V.

33. If X falls exactly on the border between two intervals, we add a half of a value to each of them. E. V.

34. The values of a continuous RV do not have any arithmetic mean, at least not in the usual sense.

35. This is wrong. Random errors of measurements are not always normal, to say nothing about the unavoidable systematic errors.

36. The author introduced both some reasonable boundaries and interval (0, 1), – perhaps coinciding with those boundaries. The Poisson distribution, see below, is applied when the studied probability is either very low or very high (a case mentioned above).

Bibliography

Chaitin G. J. (1975), Randomness and mathematical proof. *Scient. American*, vol. 232, May, pp. 47 – 52

De Morgan A. (1864), On the theory of errors of observation. *Trans. Cambr. Phil. Soc*., vol. 10, pp. 409 – 427.

De Morgan Sophia Elizabeth (1882), *Memoir of Augustus De Morgan*. London.

Feller W. (1950), *Introduction to Probability Theory and Its Applications*, vol. 1. Russian translation: Moscow, 1964.

Markov A. A. (1924), *Ischislenie Veroiatnostei* (Calculus of Probability). Moscow. Fourth edition. German edition: Leipzig – Berlin, 1912.

Pearson K. (1900), On a criterion that a given system of deviations etc. *London, Edinb. and Dublin Phil. Mag.*, ser. 5, vol. 50, pp. 157 – 175.

Poisson S.-D. (1837), *Recherches sur la probabilité des jugements* etc. Paris, 2003. English translation: **S, G,** Document No. 53.

Oscar Sheynin

Theory of Probability

An Elementary Treatise against a Historical Background

О. Б. Шейнин, *Теория вероятностей. Элементарное руководство с привлечением сведений из истории этой науки.*
Берлин, 2008

Contents

0. Introduction

0.1. Some Explanation

This treatise is written on an elementary level; in more difficult cases the final formulas are provided without proof. Nevertheless, it was impossible to leave out integrals and I also had to differentiate an integral with respect to a parameter. I include many examples taken from the history of probability and hope that my subject has thus become lively. I especially quote Karl Pearson's (1978, p. 1) repentant confession:

I do feel how wrongful it was to work for so many years at statistics and neglect its history.

In spite of a few mistakes, his book deserves serious attention. Thus, in § 4.1.1 I criticize his opinion about Jakob Bernoulli.

I have devoted much attention to the notion of probability which fully conforms to Langevin's statement (1913/1914, p. 3):

Dans toutes ces questions [in the kinetic theory] *la difficulté principale est, comme nous le verrons, de donner une définition correcte et claire de la probabilité.*

Note however that *correct definition* sounds strangely.

0.2. The Object of the Theory of Probability

Toss a coin and the outcome will be either heads or tails. Toss it 50 times and theoretically there will be 0, 1, 2, …, 49 or 50 heads. For the time being I emphasize that the number of heads will only be determined stochastically (= probabilistically): there will be from 20 to 30 heads with such-and-such probability; from 22 to 28 heads with another probability etc. In probability theory, it will never be possible to provide a quite definite answer whereas, for example, the number of roots of a given algebraic equation can be stated at once.

Games of chance (of which coin tossing is an example) was the main subject of the early theory of probability. Their outcome depends on chance rather than on the gamblers' skill, and even now they are methodically (and therefore pedagogically as well) interesting.

Many tosses provide an example of mass random events which occur in most various settings: in population statistics (births, marriages, deaths), when treating numerous observations corrupted by unavoidable random errors, applying acceptance sampling of manufactured articles with a stochastic estimation of its error, and in various branches of knowledge (kinetic theory, epidemiology etc). And so, *the theory of probability studies mass random events, or, more precisely, their regularities* which really exist. An isolated event is random, but a (homogeneous) set of events displays regularities. Aristotle (*Metaphysics* 1026b) remarked that *none of the traditional sciences busies itself about the accidental*. As stated above, neither does the theory of probability!

Laplace quite successfully applied probability to studying mass random events, and thus to investigating laws of nature (especially astronomy) and population statistics. And unlike his predecessors, he regarded the theory of probability as a branch of applied mathematics (and separated himself from the geometers: *let the geometers study* …).

I ought to add the reasonable but indefinite Laplace's opinion (1814/1886, p. CLIII): *La théorie des probabilités n'est, au fond, que le bon sens réduit au calcul*. He did not mention mass random events and furthermore his definition pertained to mathematics of his time as a whole.

Times change and we change with time … A mathematically formulated definition of the aims of the theory of probability became needed, and Boole (1851/1952, p. 251) provided it, in a seemingly dull wording: *Given the separate probabilities of any* [logical] *proposition, to find the probability of another proposition*. A similar statement pertaining to events was due to Chebyshev (1845/1951, p. 29): the theory of probability *has as its subject the determination of an event given its connection with events whose probabilities are given*. He added that probability signifies *some magnitude subject to measurement*. Prokhorov & Sevastianov (1999, p. 77) confirmed that aim and noted that such determinations were possible owing to the stability of those same mass random *phenomena*, as they stated. Anyway, owing to the stability of statistical probability (§ 1.1.3).

Since the theory of probability is axiomatized, it belongs to pure mathematics rather than a branch of applied mathematics (Laplace, see above).

Chapter 1. Main Notions, Theorems and Formulas

1.1. Probability

1.1.1. *Theoretical Probability*. Suppose that the outcome of a trial depends on n incompatible and equally possible cases only m of which are favourable for the appearance of some event A. Then its probability is assumed as

$$P(A) = m/n \qquad (1.1)$$

and it can change from 0 to 1, from an impossible to a certain event. This is the so-called classical definition due (not to Laplace, but) to De Moivre (1711/1984, p. 237) although he formulated it in the language of chances as he also did later, in 1756.

That definition had been known or intuitively applied from antiquity. The Talmud recognized seven *levels* of food containing differing relative amounts of a prohibited element (Rabinovitch 1973, p. 41). In the 14th century, Oresme (1966, p. 247) possibly thought about probability in the modern way since he stated that *two* [randomly chosen] *numbers were probably incommensurable*. (For us, his understanding of incommensurability was unusual.) The same idea of probability is seen in Kepler (1596). Finally, I cite Jakob Bernoulli (1713, Chapter 1 in Pt. 4). He introduced probability just as De Moivre did but not formally, nor did he apply it in the sequel.

In ancient times, geometry started by definitions of a point, a line and a plane. The point, for example, was something dimensionless. Nowadays, such *negative* definitions are unacceptable; just consider: a man is not a woman … and a woman is not a man! We have to accept such initial notions without defining them. Here is Markov (1900; 1908, p. 2; 1924, p. 2; and 1911/1981, p. 149):

Various concepts are defined not by words, each of which can in turn demand definition, but rather by [our] *attitude towards them ascertained little by little*.

I shall not defend these basic theorems linked to the basic notions of the calculus of probability, notions of equal probability, of independence of events, and so on, since I know that one can argue endlessly about the basic principles even of such a precise science as geometry.

Then, Kamke (1933, p. 14) noted: *Um das Jahr 1910 konnte man in Göttingen das Bonmot hören*:

Die mathematische Wahrscheinlichkeit ist ein Zahl, die zwischen Null und Eins liegt und über die man sonst nicht weis.

At that tine, Göttingen was considered the mathematical world centre, but in 1934 Hilbert, who had been working there, stated that after the Jewish scholars were ousted, the university ceased to exist. Not without reason Khinchin (1961/2004, p. 396) noted that

Each author […] *invariably reasoned about equally possible and favourable chances, attempting, however, to leave this unpleasant subject as soon as possible*.

Indeed, definition (1.1) begs the question: probability depends on equal possibilities, that is, on equal probabilities. More important, it is not really a definition, but only a formula for calculating it. Just the same, the area of a square can be calculated, but the appropriate formula does not tell us the meaning of *area*. And, finally, equal possibilities exist rarely so that the application of formula (1.1) is severely restricted.

In accord with Hilbert's recommendation (1901, Problem No. 6), the contemporary theory of probability is axiomatic, but in practice statistical probability (see § 1.1.3) reigns supreme.

Example (application of theoretical probability). Apparently during 1613 – 1623 Galileo wrote a note about a game with 3 dice first published in 1718 (David 1962, pp. 65 – 66; English translation, pp. 192 – 195). He calculated the number of all the possible outcomes (therefore, indirectly, the appropriate probabilities) and compared the appearance of 9 or 12 points and 10 or 11 points (events *A* and *B*). Both *A* and *B* occurred in six ways; thus, *A* can appear when the number of points on the dice is (3, 3, 3) or (1, 4, 4 or 2, 2, 5) or (1, 2, 6 or 1, 3, 5 or 2, 3, 4), i. e. when the number of points on each die is the same; when it is only the same on two dice; and when it is different. However, the first case is realized only once, the second case, in 3 ways, and the last one, in 6 ways. Event *A* therefore appears in 25 ways whereas event *B*, according to similar considerations, in 27 ways. The total number of possible outcomes is 216, 108 for 3, 4, …, or 10 points, and again 108 for 11, 12, …, 18 points and the probabilities of *A* and *B* are 25/216 and 27/216.

This example is instructive: it shows that the cases in formula (1.1) if unequally likely can be subdivided into equally possible ones. Galileo also stated that gamblers knew that *B* was more advantageous than *A*. They could have empirically compared not 25/216 and 27/216, but 25/52 and 27/52 by only paying attention to the two studied events.

Some definitions. When two events, *A* and *B*, have occurred, we say that their product *AB* had appeared. When at least one of them has occurred, it was the appearance of their sum, $(A + B)$, and if only one (say, *A* but not *B*), then it was their difference $(A - B)$.

Example. Two chess tournaments are to be held. The probabilities of a certain future participant to win the first place (events *A* and *B*) are somehow known. If the tournaments will occur at the same time, the product *AB* is senseless, formula (1.1) cannot be applied, the probability of that product does not exist.

1.1.1.-1. *The addition theorem.* For incompatible events *A* and *B*

$$P(A + B) = P(A) + P(B).$$

Examples. Suppose that an urn contains *n* balls, *a* of them red, *b*, blue, and *c*, white. Required is the probability of drawing a coloured ball (Rumshitsky 1966). The answer is obviously $a/n + b/n$.

Here, however, is only a seemingly similar problem. A die is rolled twice. Required is the probability that one six will appear. Call the occurrence of 6 points in the first and the second trial by *A* and *B*. Then

$$P(A) = 1/6,\ P(B) = 1/6,\ P(A + B) = 1/3.$$

But something is wrong! After 6 trials the probability will be unity, and in 7 trials?.. The point is, that *A* and *B* are not incompatible. See the correct solution in § 1.1.1-2.

The addition and the multiplication (see below) theorems for intuitively understood probabilities have actually been applied even in antiquity. Aristotle (*De Caelo* 292a30 and 289b22) stated that

Ten thousand Coan throws [whatever that meant] *in succession with the dice are impossible and it is therefore difficult to conceive that the pace of each star should be exactly proportioned to the size of its circle*.

Imagined games of chance had illustrated impossible events: the stars do not rotate around the sky randomly. Note that the naked eye sees about six thousand stars.

1.1.1-2. *Generalization: the formula of inclusion and exclusion.* For two events *A* and *B* the general addition formula is

$$P(A + B) = P(A) + P(B) - P(AB).$$

Indeed, in the example in § 1.1.1-1 $P(AB) = 1/36$ and

$$P(A + B) = 1/6 + 1/6 - 1/36 = 11/36.$$

The number of favourable cases was there 11 rather than 12. For 3 events we have

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

and in the general case

$$P(A_1A_2 \dots A_n) = P(\sum_i A_i) - P(\sum_{i<j} A_iA_j) + P(\sum_{i<j<k} A_iA_jA_k) - \dots$$

This *formula of inclusion and exclusion* was applied by Montmort (1708). It is a particular case of the proposition about the mutual arrangement of arbitrary sets. The conditions $i < j$, $i < j < k$, … ensure the inclusion of all subscripts without repetition. Thus, for 4 events $i < j$ means that allowed are events with subscripts 1, 2; 1, 3; 1, 4; 2, 3; 2, 4 and 3, 4, six combinations in all (of 4 elements taken 2 at a time).

1.1.1-3. *The multiplication theorem.* We introduce notation $P(B/A)$, denoting the probability of event B given that event A had occurred. Now, the theorem:

$$P(AB) = P(A)P(B/A). \quad (1.2)$$

Switch A with B, then

$$P(AB) = P(B)P(A/B). \quad (1.3)$$

Example 1 (Rumshitsky 1966). There are 4% defective articles in a batch; among the others 75% are of the best quality. Required is the probability that a randomly chosen article will be of the best quality.

Denote the extraction of a standard article by A, and by B, of one of the best. Then

$$P(A) = 1 - 0.04 = 0.96;\ P(B/A) = 0.75.\ P(AB) = 0.96 \cdot 0.75 = 0.72.$$

Example 2. What number of points, 11 or 12, will occur more probably in a cast of two dice? Leibniz (Todhunter 1865, p. 48) thought that both outcomes were equally probable since each was realized in only one way, when casting 5 and 6 and 6 and 6 respectively. An elementary mistake committed by a great man! Denote by A and B the occurrence of 5 and 6 on a die, then $P(A) = 1/6$, $P(B) = 1/6$. Yes, both outcomes after casting both dice are the same, $P(AB) = P(A)P(B) = 1/36$, but we ought to take into account that the first alternative can appear in two ways, 5 and 6, and 6 and 5, and is therefore twice as probable.

In general, if $P(B/A) = P(B)$ the multiplication theorem is written as

$$P(AB) = P(A)P(B),$$

and the events *A and B are called independent.*

Example 3. *A* and *B* have 12 counters each and play with 3 dice. When 11 points appear, *A* gives *B* a counter and *B* gives a counter to *A* when 14 points occur. Required are the gamblers' chances of winning all the counters. This is Additional problem No. 5 formulated by Pascal, then by Huygens (1657), who provided the answer without solution.

There are 216 outcomes of a cast of 3 dies, 15 of them favouring the appearance of 14 points, and 27 favouring 11 points, see Example in § 1.1.1. The probabilities or chances of winning are therefore as 15/27 = 5/9. For winning 12 counters the chances therefore are as $5^{12}/9^{12}$.

This was the first of a series of problems describing the *gambler's ruin*. They proved extremely interesting and among their investigators were De Moivre and Laplace. In a particular case, the fortune of one of the gamblers was supposed to be infinite.

A series of games of chance can be thought of as a random walk whereas, when considered in a generalized sense, they become a random process (§ 5.2).

Suppose now that more than 2 (for example, 4) events are studied. The multiplication theorem will then be generalized:

$$P(A_1A_2A_3A_4) = P(A_1)P(A_2/A_1)P(A_3/A_1A_2)P(A_4/A_1A_2A_3).$$

The last multiplier, for example, denotes the probability of event A_4 given that all the other events had happened.

Reader! Bear with me for some time yet; two more statements are needed, perhaps not very elegant (every man to his taste).

1.1.1-4. *A more essential generalization of the multiplication theorem.* Suppose that event *A* can occur with one and only one of several incompatible events $B_1, B_2, \ldots, B_n$. It follows that our notation $P(AB)$ can now be replaced simply by $P(A)$, so that formula (1.3) will be

$$P(A) = P(B_1)\ P(A/B_1) + P(B_2)\ P(A/B_2) + \ldots + P(B_n)\ P(A/B_n) =$$

$$\sum_{i=1}^{n} P(B_i)P(A/B_i)\,. \qquad (1.4)$$

This is the formula of *total probability* and the B_i's may be considered the *causes* of the occurrence of *A*, each leading to *A* although only with its own probability.

Suppose that 3 urns have, respectively, 1 white (w) and 2 black (b) balls; 2 w and 1 b ball; and 3 w and 5 b balls. An urn is randomly selected and a ball is drawn from it. Required is the probability that that ball is white.

The probabilities of extracting a white ball from those urns are $P(A/B_i)$ = 1/3, 2/3 и 3/8, and the probability of selecting any urn is the same, $P(B_i)$ = 1/3. Therefore,

$$P(A) = (1/3)\cdot(1/3) + (1/3)\cdot(2/3) + (1/3)\cdot(3/8) = 0.458 < 1.$$

It is quite possible that the extracted ball was black. But can $P(A) > 1$? Let

$$P(A/B_1) = P(A/B_2) = P(A/B_3) = 0.99$$

and of course

$$P(B_1) + P(B_2) + P(B_3) = 1.$$

But the feared event will not occur even when the first 3 probabilities are so high. But can $P(A) = 1$?

1.1.1-5. *The Bayes formula.* The left sides of equations (1.2) and (1.3) coincide, and their right sides are equal to each other

$$P(A)P(B/A) = P(B)P(A/B),$$

or, in previous notation,

$$P(A)P(B_i/A) = P(B_i)P(A/B_i),$$

$$P(B_i/A) = \frac{P(B_i)P(A/B_i)}{P(A)}. \qquad (1.5)$$

Replace finally $P(A)$ according to formula (1.4):

$$P(B_i/A) = \frac{P(B_i)P(A/B_i)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)}. \qquad (1.6)$$

It is time to contemplate. We assigned probabilities $P(B_1)$, $P(B_2)$, …, $P(B_n)$ to causes B_1, B_2, …, B_n and they are in the right side of (1.6). But they are prior whereas the trial was made: the event A has occurred and those prior probabilities can now be corrected, replaced by posterior probabilities $P(B_1/A)$, $P(B_2/A)$, …, $P(B_n/A)$.

Bayes (1764) included formula (1.6) but only in the particular case of $n = 1$ (which means going back to the previous formula). However, it is traditionally called after him. More precisely, from 1830 onwards it was formula (4.5) that was called after him. Nevertheless, Cournot (1843, § 88), although hesitantly, attributed formula (1.6) to Bayes; actually, it appeared in Laplace's great treatise (1812, § 26).

And who was Bayes? A talented mathematician. His posthumous memoir (1764 – 1765) became lively discussed in the early 20th century since prior probabilities were rarely known; is it possible to suppose that they are equal to each other? Laplace (1814/1995, p. 116) thought that hypotheses should be created without attributing them *any reality* and *continually* corrected by new observations. Discussions are still continuing and anyway several terms are called after Bayes, for example, Bayesian approach, estimator etc.

Example. Consider the same three urns as above. For them, the fractions in the right side of formula (1.5) differ one from another only by multipliers $P(A/B_i)$, which are to each other as (1/3):(2/3):(3/8) = 8:16:9. The same can therefore be stated about the posterior probabilities $P(B_i/A)$. It is certainly possible to take into consideration the previously established value $P(A)$ = 0. 458 and calculate them:

$$P(B_1/A) = \frac{1}{3 \cdot 3 \cdot 0.458},\ P(B_2/A) = \frac{2}{3 \cdot 3 \cdot 0.458},\ P(B_3/A) = \frac{3}{3 \cdot 8 \cdot 0.458}.$$

Understandably, probability $P(B_2/A)$ turned out as the highest of them: the relative number of white balls was largest in that same urn, the second one.

Stigler (1983/1999) applied the *Bayes theorem* in the mentioned particular case for stating that another English mathematician, Saunderson, was the real author of the Bayes memoir. He (p. 300) assigned subjective probabilities to three differing assumptions (for example, did each of them, Bayes and Saunderson, keep in touch with De Moivre) and multiplied these probabilities for each of the two. Their ratio occurred to be 3:1 in favour of the latter. Tacitly allowing an equality of the corresponding prior probabilities, Stigler (p. 301) decided that the probability of Saunderson's authorship was three times higher. Stigler's tacit assumption was absolutely inadmissible and that his (happily forgotten) conclusion ought to be resolutely rejected. That same Stigler allowed himself to vomit an abuse on Euler (§ 6.2) and Gauss (Sheynin 1999a, pp. 463 – 466).

1.1.1-6. *Subjective probability*. It is naturally prior and somewhat complements the theoretical probability (1.1). Indirectly, it is applied very often, especially when there exists *no reason* to doubt the existence of equal probabilities of some outcomes. Thus, the probability of each outcome of a cast of a die is supposed to be 1/6, although any given die is not exactly *regular*. Poisson and Cournot (1843/1984, p. 6) were the first to mention it definitely. They even called it and the objective probability by different terms, *chance* and *probability*.

Here is Poisson's (1837, § 11) instructive problem. An urn contains n white and black balls in an unknown proportion. Required is the probability that an extracted ball is white. The number of white balls can be 0, 1, 2, …, n, – (n + 1) allegedly equally probable cases. The probability sought is therefore the mean of all possible probabilities

$$\frac{1}{n+1}\left(\frac{n}{n} + \frac{n-1}{n} + \ldots + \frac{1}{n} + \frac{0}{n}\right) = \frac{1}{2}$$

as it should have been. His answer conforms to the principles of the information theory which Poisson himself understood perfectly well: it, his answer, corresponded to *la perfaite perplexité de notre esprit*.

Poisson (1825 – 1826) applied subjective probability when investigating a game of chance. Cards are extracted one by one from six decks shuffled together as a single whole until the sum of the points in the sample obtained was in the interval [31; 40]. The sample is not returned and a second sample of the same kind is made. It is required to determine the probability that the sums of the points are equal. Like the gamblers and bankers, Poisson tacitly assumed that the second sample was extracted as though from the six initial fresh decks. Actually, this was wrong, but the gamblers thought that, since they did not know what happened to the initial decks, the probability of drawing some number of points did not change.

When blackjack is played, bankers are duty bound to act the same wrong way: after each round the game continues without the used cards, and, to be on the safe side, they ought to stop at 17 points. A gambler endowed with a retentive memory can certainly profit from this restriction.

Here are other examples. *Redemption of the first born*. The Jerusalem Talmud (Sanhedrin 1[4]) describes how lots were taken. The main point was that the voters were afraid that there will be no more special ballots left freeing the last voters from the payment. They actually thought about the subjective probabilities of the distribution of those special ballots among consecutive voters. Tutubalin (1972, p. 12) considered the same problem in quite another setting and proved that the fears of the voters were unfounded.

Another example. Rabinovitch (1973, p. 40) desribed the statement of Rabbi Shlomo ben Adret (1235 – 1310 or 1319) about eating some pieces of meat one of which was not kosher. One piece after another may be eaten because (actually) the probability of choosing the forbidden piece was low, and when only two pieces are left, – why, the forbidden piece was already eaten, so eat these two as well!

Subjective *opinions* are mathematically studied, for example those pertaining to expert estimates and systems of voting. In those cases the merits of the economic projects or candidates are arranged in ascending or descending order of preference, see § 5.1.

1.1.2. *Geometrical Probability.* The classical definition of probability can be generalized, and, in a manuscript of 1664 – 1666, Newton (1967, pp. 58 – 61) was the first to do so. He considered a ball falling upon the centre of a circle divided into sectors whose areas were in *such proportion as* 2 *to* √5. If the ball *tumbles* into the first sector, a person gets *a*, otherwise he receives *b*, and his *hopes is worth*

$$(2a+b\sqrt{5})\div(2+\sqrt{5}).$$

The probabilities of the ball tumbling into these sectors were as 2 to √5, as Newton also indirectly stated. See also Sheynin (1971a).

The classical definition is still with us with *m* and *n* being real rather than only natural numbers. In this way many authors effectively applied *geometrical probability*. Buffon (1777, § 23) definitively introduced it by solving his celebrated problem. A needle of length 2*r* falls *randomly* on a set of parallel lines. Determine the probability P that it intersects one of them. It is easily seen that

$$P = 4r/\pi a$$

where $a > 2r$ is the distance between adjacent lines. Buffon himself had however only determined the ratio r/a for $P = 1/2$. His main aim was to *mettre donc la Géométrie en possession de ses droits sur la science du hazard* (Buffon 1777/1954, p. 471). Later authors generalized the Buffon problem, for example by replacing lines by rectangles or squares.

Laplace (1812, chapter 5) noted that after, say, 100 such trials the number π can be calculated. He thus suggested the Monte Carlo method (of statistical simulation). A formal definition of the new concept was only due to Cournot (1843, § 18). More precisely, he offered a general definition for a discrete and a continuous random variable by stating that probability was the ratio of the *étendue* of the favourable cases to that of all the cases. We would now replace *étendue* by *measure* (in particular, by area).

Actually, beginning with Nikolaus Bernoulli (1709/1975, pp. 296 – 297), see also Todhunter (1865, pp. 195 – 196), each author dealing with continuous laws of distribution (§ 2.1) applied geometric probability. The same can be said about Boltzmann (1868/1909, p. 49) who defined the probability of a system being in a certain phase as the ratio of the time during which it is in that time to the whole time of the motion. Ergodic theorems can be mentioned, but they are beyond our boundaries.

Determine the probability that a random chord of a given circle is shorter than the side of an inscribed equilateral triangle (Bertrand 1888). This celebrated problem had been discussed for more than a century and several versions of *randomness* were studied. Bertrand himself offered three different solutions, and it was finally found out that, first, there was an uncountable number of solutions, and, second, that the proper solution was *probability equals* 1/2 which corresponded to *la perfaite perplexité de notre esprit* (§ 1.1.1-6).

Finally, the encounter problem (Laurent 1873, pp. 67 – 69): two persons are to meet at a definite spot during a specified time interval (say, an hour). Their arrivals are independent and occur *at random*; the first one to come waits only for a certain time (say, 20 minutes), then leaves. Required is the probability of a successive encounter.

Denote the time of their arrivals by x and y, then $|x - y| \leq 20$ or $|y - x| \leq 20$, and a graphical solution is simple and instructive, see also § 3.2.

1.1.3. *Statistical Probability.* Suppose that a random event occurred μ times in ν trials. Then its relative frequency (frequency, as I will call it) or statistical probability is

$$\hat{p} = \mu/\nu \qquad (1.7)$$

and it obviously changes from 0 to 1.

Newton (§ 1.1.2), while commenting on his second thought experiment, a roll of an irregular die, concluded that, nevertheless, *It may be found how much one cast is more easily gotten then another*. He likely had in mind statistical probabilities rather than analytic calculations. And he may well have seen Graunt's pioneer statistical contribution of 1662 where all deductions pertaining to population and medical statistics had been based on statistical probabilities.

Statistical probability was applied even by Celsus (1935, p. 19) in the first century of our era:

Careful men noted what generally answered the better, and then began to prescribe the same for their patients. Thus sprang up the Art of medicine.

He certainly had no numerical data at his disposal, but qualitative statements had been a distinctive feature of ancient science.

The definition above is only meaningful if the trials are mutually independent and the calculated probability remains almost the same in a subsequent series of similar trials. If results of some trials essentially differ, say, from one day of the week to another, then each such day ought to be studied separately. And what kind of trials do we call independent? For the time being, we say: trials, whose results do not influence each other, also see § 1.1.1-3.

The imperfection of the theoretical probability and its narrow field of applications led to the appearance of the statistical probability as the main initial notion (Richard Mises, in the 1920s).

A rigorous implementation of his simple idea proved extremely difficult and discussions about the Mises' *frequentist theory* never ended. Here is his idea. Toss a coin many times and from time to time calculate the frequency (1.7) of heads. After a sufficiently large ν it will only change within narrow bounds and at $\nu \to \infty$ it will reach some limiting value. It was this value that Mises called statistical probability (of heads).

Infinitely long trials are impossible, but Mises cited a similar approach in physics and mechanics (for example, velocity at a given moment). He also stated that the sequence of the trials (the *collective*) should be irregular (so that its infinite subsequences should lead to the same probability $\hat{p}$).

This condition is too indefinite. How many subsequences ought to be tested before irregularity is confirmed? And is it impossible to select *randomly* an excessively peculiar subsequence? Even these superficial remarks show the great difficulties encountered by the frequentist theory; nevertheless, naturalists have to issue from statistical probability.

Yes, it is theoretically imperfect, although mathematicians came to regard it somewhat milder (Kolmogorov 1963, p. 369). I ought to add that (Uspensky et al 1990, § 1.3.4)

Until now, it proved impossible to embody Mises' intention in a definition of randomness satisfactory from any point of view.

1.1.4. *Independence of Events and Observations.* Events A and B are independent if (1.1.1-3)

$$P(AB) = P(A)P(B),$$

otherwise

$$P(AB) = P(A)P(B/A).$$

Switch A and B, then

$$P(AB) = P(B)P(A/B).$$

A remarkable corollary follows: if A does not depend on B, then B does not depend on A; independence is mutual (De Moivre (1718/1756, p. 6):

Two events are independent, when they have no connection one with the other, and that the happening of one neither forwards nor obstructs the happening of the other.

Two events are dependent, when they are so connected together as that the probability of either's happening is altered by the happening of the other.

The proof of mutuality of independence (already evident in that definition) is simple. According to the condition, $P(A/B) = P(A)$, then by formulas (1.3) and (1.2)

$$P(AB) = P(B)P(A),\ P(B/A) = P(B), \text{ QED.}$$

Here, however, is a seemingly contradicting example. Suppose that the weather during a summer week in some town is random. Then the random sales of soft drinks there depend on it although there simply cannot be any inverse dependence. But weather and sales cannot be here considered on the same footing.

De Moivre (1711, Introduction) was the first to mention independence, see also just above. Later classics of probability theory mentioned independence of events as well (see below), but some authors forgot about it. The situation had abruptly changed since Markov investigated his *chains* (§ 5.2) and thus added an additional direction to the theory, the study of dependent random events and variables.

Gauss (1823, § 18) stated that if some observation was common to two functions of the results of observations, the errors of these latter will not be independent from each other. He added (for some reason, only in § 19) that those functions were linear. Without this restriction his statement would have contradicted the Student – Fisher theorem about the independence of the sample mean and variance in case of the normal distribution.

Also dependent, as Gauss (1828, § 3) thought, were the results of adjustments. Thus, after the observed angles of a triangle were corrected, and their sum became strictly equal to its theoretical value, these adjusted angles were not anymore independent; they are now somehow connected by their unavoidable residual errors. Note that Gauss had thus considered independence of functions of random variables (§ 1.2.3).

Geodesists invariably (and without citing Gauss) kept to the same definition. Thus, in the Soviet Union separate chains of triangulation had bases and astronomically determined directions on both ends. Therefore, after their preliminary adjustment they were included in a general adjustment as independent entities. True, the bases and those directions were common to at least two chains, but they were measured more precisely than the angles.

Bernstein (1946, p. 47) offered an instructive example showing that pairwise independence of, say, three events, is not sufficient for their mutual independence.

1.2. Randomness and Random Variables

1.2.1. *Randomness*. In antiquity, randomness was a philosophical notion, then became a mathematical concept as well. Aristotle included it in his doctrine of causes; here are two of his celebrated examples.

1) Digging a hole for a tree, someone finds a buried treasure [not a rusty nail!] (*Metaphysics* 1025a).

2) Two men known to each other meet suddenly (*Physics* 196b30); two independent chains of events *suddenly* intersected.

These examples have a common feature: a small change in the action(s) of those involved led to an essential change: the treasure would have remained buried, there would have been no meeting. Many ancient authors imagined chance just as Aristotle did whereas Cournot (1843, § 40) mentioned the second example anew.

The pattern small change – essential consequences became Poincaré's (1896/1987, pp. 4 – 6) main explanation of randomness, although he specified: when equilibrium is unstable. Here is his or, rather, Cournot's (1843, § 43) example: a right circular cone standing vertically on its vertex falls in a random direction. A similar example is due to Galen (1951, p. 202), a Roman physician and naturalist, 2nd century:

In those who are healthy […] *the body does not alter even from extreme causes; but in old men even the smallest causes produce the greatest change.*

Corruption of nature's aims was another cause of randomness. Kepler (1618 – 1621/1952, p. 932) established that planets move along elliptical orbits whereas nature, as he thought, aimed at circular orbits. Complete perfection was not attained. Only Newton proved that the ellipticity followed from his law of universal gravitation and that the eccentricity of an orbit was determined by the planet's initial velocity.

Following Kepler and Kant, Laplace (1796/1884, p. 504) somehow concluded that these eccentricities had been caused by variations of temperatures and densities of the diverse parts of the planets.

A mathematical theory cannot however be based on encounters or nature's aims. I leave aside very interesting but occurring much ahead of their time and therefore unsuccessful attempts mathematically to determine randomness (Lambert 1771, §§ 323 – 324; 1772 – 1775), see also Sheynin (1971b, pp. 245 – 246). Modern attempts deal with infinite (and even finite) sequences of zeros and unities such as

0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, …

Is it random or not? Such questions proved fundamental. They were approached in various ways, but are far from being solved. For a finite sequence that question is still more complicated. In any case, the beginning of an infinite sequence ought to be irregular so that irregularity (as Mises also thought) is an essential property of randomness.

In philosophy, randomness is opposed to necessity; in natural sciences Poincaré (1896/1912, p. 1) described their dialectic:

Dans chaque domaine, les lois précises ne décidaient pas de tout, elles traçaient seulement les limites entre lesquelles il était permis au hasard de se mouvoir.

He did not regrettably mention regularities of mass random events. It is also appropriate to recall the celebrated Laplace's (1814/1995, p. 2) statement allegedly proving that he rejected randomness:

An intelligence that, at a given instant, could comprehend all the forces by which nature is animated […], *if, moreover, it were vast enough to submit these data to analysis, would encompass* […] *the movements of the greatest bodies and those of the slightest atoms.* […] *Nothing would be uncertain, and the future, like the past, would be open to its eyes.*

Such intelligence is impossible. Then, there exist unstable motions, responding to small errors of the initial conditions (see above) and perhaps half a century ago a mighty generalization of the former phenomenon, the chaotic motion, was discovered and acknowledged. Finally, Maupertuis (1756, p. 300) and Boscovich (1758, § 385) kept to the same *Laplacean determinism*.

Allegedly proving … Perhaps Laplace's entire astronomical investigations and certainly all his statistical work refute his statements (which really took place) denying randomness.

1.2.2. *Cause or Chance*? What should we think if a coin falls on the same side 10 or 20 times in succession? Common sense will tell us: the coin was imperfect. Nevertheless, we will discuss this example. Indeed, after the appearance, in mid-19th century, of the non-Euclidean geometry we may only trust common sense in the first approximation.

Denote heads and tails by + and –. After two tosses the outcomes can be + +, + –, – + and – –, all of them equally probable. After the third toss the outcome + + becomes either + + +, or + + –. In other words, the outcome + + + is not less probable than any of the other 7, and it is easy to see that a similar conclusion remains valid at any number of tosses. Of course 10 heads in succession are unlikely, but all the other possible outcomes will be just as unlikely.

So let us refer to Laplace (1776, p. 152; 1814/1995, p. 9), who discussed the so-called D'Alembert – Laplace problem:

Suppose we laid out […] *the printer's letters Constantinople in this order. We believe that this arrangement is not due to chance, not because it is less possible than other arrangements.* […] [*S*]*ince we use this word it is incomparably more probable that someone has arranged the preceding letters in this order than that this arrangement happened by chance.*

No formulas can help us and Laplace had to turn to common sense. In our example, we may conclude that someone had done something so that the coin always falls on the same side. Common sense did not let us down. In 1776, Laplace selected the word *Infinitesimal*; it was D'Alembert (1767, pp. 245 – 255) who wrote *Constantinople*. His considerations were not as reasonable.

In general, the *cause or chance* problem compels us to separate somehow equally possible cases (if they exist) into ordinary and remarkable; *Constantinople* was indeed a remarkable arrangement. Kepler was an astrologer as well (and called himself the founder of an allegedly scientific astrology which only admitted a correlative influence of the *stars* on human beings). He (1601, § 40/1979, p. 97) added three aspects (remarkable mutual positions of the heavenly bodies) to the five recognized by the ancients and he (1604/1977, p. 337) also was *not willing to ascribe* the appearance of a New star *to blind chance* […] and considered it *a great wonder*.

Another related subject is the superstition and self-delusion peculiar to gamblers (and not only to them). A ball is rolled along a roulette wheel and stops with equal probability at any of the 37 positions 0, 1, …, 35, 36. Gamblers attempt to guess where exactly will the ball stop and the winner gets all the stakes; however, if the ball stops at 0, the stakes go the banker. This is the simplest version of the game.

Now suppose that the ball stopped at 18 three times in succession; should a gambler take this fact into account (and how exactly)?

Petty (1662/1899, vol. 1, p. 64) resolutely opposed games in chance (considered that playing as such was a superstition): *A lottery* […] *is properly a tax upon unfortunate self-conceited fools*. Montmort (1708/1980, p. 6) and other authors noted the gamblers' superstitions; and here is Laplace (1814/1995, p. 92) commenting on a similar event:

When one number has not been drawn for a long time […], *the mob is eager to bet on it*.

But it was Bertrand (1888, p. XXII) who dealt the final blow (although did not convince the gamblers): *Elle* [the roulette] *n'a ni conscience ni mémoire*. *Play, but do not retrieve your losses* (a Russian saying quoted by Pushkin)! It means: play if you cannot abstain from gambling, but never risk much. Arnauld & Nicole (1662/1992, p. 332) warned against expecting large gains (and risking much!).

Laplace (Ibidem, p. 93) also mentioned the general public' superstitions:

I have seen men, ardently longing for a son […]. *They fancied that the boys already born* [during a certain month] *made it more probable that girls would be born next*.

Finally, I note that Laplace (p. 93) saw no advantage in repeatedly staking on the same number. This brings us to martingales, but I will not go thus far.

1.2.3. *Random Variable.* This is the central notion of the theory of probability. Here is the simplest definition of a discrete random variable: *A variable taking various discrete values, each with some probability*. Denote these values by $x_1, x_2, \ldots, x_n$. The sum of their probabilities $p_1, p_2, \ldots, p_n$ should be unity. Considered together, those values and probabilities are the random variable's *law of distribution*. A random event can be understood as a random variable having $n = 2$.

The case of $n \to \infty$ is also possible; it can be realized in the discrete way, for example, if $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, …, with a *countable* number of the values, or, if a continuous random variable is considered, that number is uncountable. Example: all the uncountable values in interval [0; 1]. A new circumstance appears when there are infinitely many values: an event having a zero probability is possible. Indeed, select any point, say, in the appropriate interval. The probability of choosing any given point is certainly zero, but we did select some point! The geometric probability (§ 1.1.2) can be recalled here.

A random variable (or its generalization, which we will not discuss) *or a random event ought to be present in each problem of the theory of probability*. Thus, the outcome of a dice-fall is a random variable; it takes 6 values, each with its own probability (here, they are identical).

Many interesting examples of random variables can be provided. Thus, in the beginning of the 17th century the participants in the celebrated Genoese lottery could guess 1, 2, …, 5 numbers out of 90. The gains increased with those numbers, but the more did the gambler hope for, the heavier was he punished (his expected gain rapidly decreased). This did not follow from any mathematical theorem, but was caused by the organizers' greed.

The random variable involved (the random gain) had 5 values with definite probabilities although only a handful of people had been able to calculate them. Then, from 1662 onward (Graunt), human lifespan began to be studied. In 1756 and 1757 Simpson effectively introduced random variables into the future theory of errors and until about the 1930s this new direction of research had remained the main subject of probability theory. Simpson assumed that the chances of the (random) errors corrupting each measurement (of a given series) are represented by some numbers; the result of measurement thus became a possible value of some random variable and a similar statement held for all of them taken together.

A formal introduction of the random variable was due to Poisson (1837, pp. 140 – 141 and 254) who still called it by a certainly provisional term *chose A*. The proper term, random variable, did not come into general use all at once. Perhaps its last opponent was Markov (letter to Chuprov of 1912; Ondar 1977/1981, p. 65):

Everywhere possible, I exclude the completely undefined expression random and at random. Where it is necessary to use them, I introduce an explanation corresponding to the pertinent case.

He had not however devised anything better and often wrote *indefinite magnitude*, which was hardly better. Markov had not applied the terms *normal distribution* or *correlation coefficient* either!

In a certain sense, the entire development of the theory of probability consisted in an ever more general understanding of *random variable*. At first, randomness in general had been studied (actually, a random variable with a uniform law of distribution, see § 2.2.1) as contrary to necessity, then random variables having various distributions, dependent variables and random functions, cf. § 5.1. The level of abstraction in the theory gradually heightened (the same is true about the development of mathematics in general). It is well known that, the higher became that level (i. e., the further mathematics moved away from nature), the more useful it was. Complex numbers and functions of complex variables are absolutely alien to nature, but how useful they are in mathematics and its applications!

Chapter 2. Laws of Distribution of Random Variables, Their Characteristics and Parameters

2.1. Distribution Function and Density

For describing a continuous random variable (call it ξ) we need to determine its law of distribution as it was done in § 1.2.3 for discrete variables. Denote by $F(x)$ the probability of its being less than some x:

$$P(\xi < x) = F(x).$$

This $F(x)$ is called the distribution (integral) function of ξ. If ξ takes any value from $-\infty$ to ∞, then

$$P(\xi < -\infty) = F(-\infty) = 0, P(\xi < \infty) = F(\infty) = 1.$$

Choose now two arbitrary points, x_1 and x_2, $x_2 > x_1$, then

$$P(\xi < x_2) \geq P(\xi < x_1) \text{ or } F(x_2) \geq F(x_1).$$

Indeed, $P(-\infty < \xi < x_2)$ cannot be lower than $P(-\infty < \xi < x_1)$. And if a random variable takes no values on interval $[x_1; x_2]$ (but remains continuous beyond it), then

$$P(\xi < x_2) = P(\xi < x_1) \text{ or } F(x_2) = F(x_1). \quad (2.1)$$

And so, in any case, the function $F(x)$ does not decrease and if (2.1) does not take place, it increases. Note also that

$$F(x_2) - F(x_1) = P(\xi < x_2) - P(\xi < x_1). \quad (2.2)$$

Integral distribution functions began to be applied in the 20th century, although they fleetingly appeared even in 1669. Pursuing a methodical aim, Huygens (1669/1895, between pp. 530 и 531) drew a graph of a function whose equation can be written as

$$y = 1 - F(x), 0 \leq x \leq 100.$$

The curve described the human lifespan (ξ), the probability of $P(\xi \geq x)$, but it was not based on numerical data. In 1725, De Moivre studied the same probability, and similarly Clausius (1858/1867, p. 268) investigated the probability of the free path of a molecule to be not less than x.

Until distribution functions really entered probability, continuous random variables had been described by *densities* $\varphi(x)$ of their distributions (of their probability). Consider an infinitely short interval

[x_1; $x_1 + dx_1$]. A random variable takes there a value depending on x_1; we may say, takes one and the same value $\varphi(x_1)$. On the adjacent interval of the same length on the right side the value of that variable may be assumed equal to $\varphi(x_2)$, $x_2 = x_1 + dx_1$. Thus we get a series of values $\varphi(x_1)$, $\varphi(x_2)$, … and can describe the relation of this function, $\varphi(x)$, the density, with $F(x)$:

$$F(x_n) = \int_{-\infty}^{x_n} \varphi(x)dx,\; F(x_1) = \int_{-\infty}^{x_1} \varphi(x)dx,\; F(x_n) - F(x_1) = \int_{x_1}^{x_n} \varphi(x)dx\,.$$

These formulas additionally explain equality (2.2). Strictly speaking, by definition,

$$F'(x) = \varphi(x),$$

but the essence of $\varphi(x)$ as stated above certainly holds. In more simple examples the density is a continuous function existing on a finite or infinite interval; according to its definition, the area *under* the density curve is unity.

Under, above, to the left or to the right are non-mathematical expressions, but we will apply them even without italics.

Instead of random variables themselves the theory of probability studies their distribution functions or densities just as trinomials

$$f(x) = ax^2 + bx + c,\ a \neq 0$$

are studied in algebra. Given the parameters a, b and c, we can determine whether the roots of the trinomial are real (coinciding or not) or complex, can draw its graph. The same way we determine the behaviour of random variables. But where are the parameters of densities or distribution functions?

Consider a function $f(x)$. We may write it down as $f(x; a; b; c)$ and thus show that its argument is the variable x, but that its behaviour is also determined by parameters constant for each given function (for each trinomial). The density and the distribution function also have parameters peculiar to each random variable. As a rule, statisticians estimate those parameters. Suppose that we have a continuous triangular distribution (assumptions of such kind should be justified) with an unknown parameter a (see § 2.2.2). It is required to *estimate* it, to establish for it some (sample) value $\hat{a}$, which is only possible when having appropriate observations of the random variable, and to determine the possible error of that estimate. If there are two parameters, certainly both should be estimated.

2.2. Some Distributions

2.2.1. *The uniform distribution*. A random variable having this distribution takes all its values with the same probability. Thus, all the 6 outcomes of a die-fall are equally probable. A continuous random variable takes identical values on some interval. The area under this interval should be unity; for interval [– a, a] the density will therefore be

$$\varphi(x) = 1/a = \text{Const}$$

and a can be considered the parameter of this distribution.

2.2.2. *The continuous triangular distribution* is usually even. So let it cut the x-axis at points A $(-a, 0)$ and C $(a, 0)$. The density is the broken line ABC with AB and BC being the equal lateral sides of the isosceles triangle ABC. The area under it is unity, so we have B$(0, 1/a)$.

The only parameter of this distribution is obviously a since only it determines the coordinates of all the points A, B and C. I described the triangular distribution mostly since it was easy to establish the meaning of its parameter. It was introduced by Simpson (§ 1.2.3).

2.2.3. *The binomial distribution.* We all remember the formula of the square of the sum of two numbers and some of us even managed to remember the formula for the cube of the same sum. However, there exists a general formula for natural exponents $n = 1, 2, 3, \ldots$:

$$(p + q)^n = p^n + C_n^1 p^{n-1}q + C_n^2 p^{n-2}q^2 + \ldots + C_n^{n-1} pq^{n-1} + q^n. \quad (2.3)$$

We are only interested in the particular case of $p + q = 1$, that is, in those magnitudes which describe the probabilities of contrary events. Here, C_n^k is the number of combinations of n taken k at a time:

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

The numerator has the same number of multipliers as the denominator. Thus,

$$C_5^3 = \frac{5\cdot 4\cdot 3}{3!},\ 3! = 1\cdot 2\cdot 3.$$

Required now is the probability of casting a unity twice when rolling four dice (or rolling one die four times). Cast a die once, and the probability of a unity is $p = 1/6$, whereas the probability of all the other outcomes is $q = 5/6$. And now consider a binomial $[(1/6) + (5/6)]$ raised to the fourth power:

$$[(1/6) + (5/6)]^4 = [1/6^4](1 + 4\cdot 1^3\cdot 5 + 6\cdot 1^2\cdot 5^2 + 4\cdot 1\cdot 5^3 + 5^4).$$

The term $6\cdot 1^2\cdot 5^2$ will correspond to the probability sought since it, and only it, includes the multiplier 1^2, denoting the studied outcome (and another outcome). That probability is $6[1/6^4]\cdot 1^2\cdot 5^2 = 25/6^3 = 25/216$. We have thus taken into account the number (6) of the possible successive casts (the number of combinations of 4 elements taken 2 at a time). Neglecting this coefficient 6, we would have obtained the probability sought when the successful casts were fixed; for example, if the unity should have occurred in the first and the third roll.

The number of trials *n* and the ratio *p/q* can be chosen as the parameters of the binomial distribution (2.3). It is not necessary to choose both *p* and *q* since only one of these magnitudes is independent ($p + q = 1$). The example above shows that each term of the binomial expansion (2.3) is the probability

$$p(x) = C_n^k p^{n-k} q^k,\ x = 0, 1, 2, \ldots, n$$

that the studied random event will occur *k* times in whichever *n* trials. The frequency is also essential, see § 2.4.1.

Interesting examples of the binomial distribution include the studies of the sex ratio at births, cf. § 4.2. Its generalization is the multinomial distribution with each trial concerning a random variable taking several values rather than a random event. It is therefore described by a multinomial

$$(a + b + c + \ldots)^n.$$

Pertinent qualitative reasoning without mentioning probabilities were due to Maimonides (Rabinovitch 1973, c. 74):

Among contingent things some are very likely, other possibilities are very remote, and yet others are intermediate.

2.2.4. *The normal distribution.* The function

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-a)^2}{2\sigma^2}],\ -\infty < x < \infty, \qquad (2.4)$$

is the density of the normal distribution. The stochastic meaning of the two of its parameters, *a* and $\sigma > 0$, is described in § 2.4.2. The corresponding distribution function is

$$F(z) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{z}\exp[-\frac{(x-a)^2}{2\sigma^2}]dx.$$

Let $a = 0$ and $\sigma = 1$, then, in the *standard* case,

$$F(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp[-\frac{x^2}{2}]dx. \qquad (2.5)$$

It is however more convenient to tabulate the function

$$F(z) = \frac{1}{\sqrt{2\pi}}\int_{0}^{z}\exp[-\frac{x^2}{2}]dx. \qquad (2.6)$$

Indeed, the integrand in formula (2.5) is an even function so that the integrals (2.5) within (– ∞; 0] and [0; + ∞), are equal to each other and equal to 1/2; within, say, (– ∞; – 1] the integral (2.5) is equal to the difference between 1/2 and integral (2.6) at $z = 1$. The value of the function (2.6) at $z \approx 3$ is already 0.499; if $z \to +\infty$ its value is 1/2, or, which is the same, within infinite limits its value is unity, as it should be.

The utmost importance of the normal distribution follows from the so-called *central limit theorem* (CLT), a term due to Polya (1920):

The sum of a large number of independent random variables, each of them only to a small degree influencing that sum, is distributed normally.

It was Pearson, who, in 1893, definitively introduced the term *normal distribution* in order to avoid naming it after Gauss (1809) or Laplace who extensively applied it after non-rigorously proving several versions of the CLT. Galton applied that term before Pearson, but the first to suggest it was Peirce (1873, p. 206).

De Moivre (§ 4.2) considered the appearance of the normal law from a binomial distribution and thus proved a particular case of the CLT. Many authors not to mention Laplace had proved various versions of the CLT, but its rigorous proof was due to Markov and Liapunov, not even to Chebyshev.

Denote the probabilities of a male and female births by p and q and neglect all the other possible births so that $p + q = 1$. Then the probabilities of some number of male births (or of this number remaining within some bounds) can be calculated by means of the normal distribution. This was indeed De Moivre's immediate aim. From 1711 onward the parameter p/q became an object of numerous studies (§ 4.2).

About 1874 Galton (1877) invented the so-called *quincunx*, a device for visually demonstrating the appearance of the normal distribution as the limiting case of the uniform law. Shot was poured through several (say, 20) lines of pins, and each shot 20 times deviated with the same probability to the right or to the left and finally fell on the floor of the device. Thus appeared a normal curve. A special feature of that device was that it showed that the normal law was stable (§ 6.1).

2.2.5. *The Poisson distribution.* The law of this discrete distribution (Poisson 1837, p. 205) can be written down as

$$P(x) = \frac{a^x}{x!}e^{-a},\ x = 0, 1, 2, \ldots$$

The sum of the probabilities $P(x)$ over all the infinite set of the values of x is 1, as it should be. Indeed, e^{-a} is the common multiplier and

$$\sum_{x=0}^{\infty}\frac{a^x}{x!}e^{-a} = e^{-a}(1+\frac{a}{1!}+\frac{a^2}{2!}+\frac{a^3}{3!}+\ldots) = e^{-a}e^{a} = 1.$$

Here is an interesting pattern leading to the Poisson distribution: points are entered on an interval according to a uniform distribution, one by one, independently from each other. It occurs that the number of points situated on some part of that interval obeys the Poisson distribution. Example: the number of calls entering an exchange. Its functioning can therefore be stochastically studied.

Suppose an exchange serves 300 subscribers and the hourly probability of one of them speaking is $p = 0.01$. What will be the probability of four or more independent calls made during an hour? The conditions for the appearance of the Poisson distribution are met, and $a = pn = 3$. Then

$$P(\xi \geq 4) = \sum_{x=0}^{\infty} \frac{a^x}{x!} e^{-a} - P(\xi = 0) - P(\xi = 1) - P(\xi = 2) - P(\xi = 3).$$

The sum is unity (see above) and the other terms are easily calculated.

Another example: the distribution of the stars over the sky (Michell 1767). If they are distributed uniformly (on a sphere rather than interval), some of them will be very close to each other (double, triple, … stars). Even then many such stars had been known, and Michell questioned whether this occurred randomly or not. What is the probability that two stars out of all of them are situated not more than 1° apart?

Newcomb (1860, pp. 427 – 429) applied the Poisson distribution to derive the probability that some small part of the celestial sphere contains s stars out of n uniformly distributed across the celestial sphere. In a sense, it is this distribution that best describes a *random* arrangement of many points. Its parameter is obviously a.

In 1898 Bortkiewicz introduced his *law of small numbers*, and for a few decades it had been considered as the main law of statistics. Actually, it only popularized the then yet little known Poisson distribution which is what Kolmogorov (1954) stated but did not justify his opinion and I (2008) proved that he was correct. Botkiewicz's contribution is deservedly forgotten although mostly owing to previous more particular criticisms.

2.2.6. *The hypergeometric distribution*. It is important for acceptance inspection of mass production, see below. Consider the Additional problem No. 4 (Huygens 1657) first formulated by Pascal. Given, 12 counters, 4 of them white (as though defective). Required is the probability that 3 white counters occur among 7 counters drawn without replacement.

Well, actually the entire batch should be rejected, but nevertheless I go ahead following Jakob Bernoulli (1713, part 3, problem 6), although applying the hypergeometric distribution. Huygens, it ought to be added, provided the answer, but not the solution. Denote the conditions of the problem: $N = 12$, $M = 4$, $n = 7$, $m = 3$. Simple combinatorial reasoning lead to a formula which is indeed the formula of that distribution:

$$P(\xi = m) = C_M^m C_{N-M}^{n-m} \div C_N^n.$$

2.3. The Main Characteristics of Distributions

2.3.1. *Expectation*. For a discrete random variable ξ it is the sum of the products of all its values $x_1, x_2, …, x_n$ by their probabilities $p_1, p_2, …, p_n$:

$$\text{E}\xi = \frac{p_1x_1 + p_2x_2 + ... + p_nx_n}{p_1 + p_2 + ... + p_n}. \qquad (2.7)$$

The denominator is naturally unity. Laplace (1812/1886, p. 189) added the adjective *mathematical* to expectation so as to distinguish it from the then topical but now forgotten moral expectation (see below). This adjective is regrettably still applied in French and Russian literature.

Expectation can be considered a natural ersatz of a random variable, as though its mean value; in the theory of errors, it corresponds to the generalized arithmetic mean. Denote observations by $x_1, x_2, …, x_n$, and their weights (worth) by $p_1, p_2, …, p_n$. By definition their mean is

$$\overline{x} = \frac{p_1x_1 + p_2x_2 + ... + p_nx_n}{p_1 + p_2 + ... + p_n}, \qquad (2.8a)$$

although the denominator is not 1 anymore. If all the weights are identical

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}. \qquad (2.8b)$$

In § 2.6 I mentioned the selection of bounds covering a measured constant as practised by ancient astronomers. Here, I note that they did not choose any definite estimator, such as the arithmetic mean; they had applied qualitative considerations and thought about convenience of subsequent calculations. For observations corrupted by large errors this tradition makes sense.

So when had that mean become the standard estimator? While selecting a mean of four observations, Kepler (1609/1992, p. 200/63) chose a generalized mean (2.8a) rather than *the letter of the law*, i. e., as I understand him, rather than the mean (2.8b), see Sheynin (1993b, p. 186).

The mean (2.8a) had sometimes been applied with posterior weights p_i, equally decreasing on either side of the middle portion of the observations. This choice is hardly useful since, first, these weights are necessarily subjective; and, second, since that estimator only provided a correction of the mean (2.8a) for the unevenness of the sample density of probability of the observational errors.

The expectation (2.7) and the arithmetic mean (2.8) nevertheless essentially differ. The former is a number since it presumably contains all the values of a random variable, whereas the latter is compiled from the results of observations unavoidably corrupted by random errors (as well as by systematic errors, but now we do not need them) and is therefore a random variable as well, as though a sample value of the unknown expectation. Its error ought to be estimated and *a similar remark will also apply to other characteristics of a random variable*.

At the same time the arithmetic mean is assumed as the value of the measured constant (§ 6.2). Note that notation $\overline{x}$ for the values of x_i is standard.

For a continuous random variable the expectation is expressed by the integral

$$\mathrm{E}\xi = \int_a^b x\varphi(x)dx. \qquad (2.9)$$

Points a and b are the extreme points of the domain of the density $\varphi(x)$ and possibly $a = -\infty$, and $b = \infty$.

Expectation had begun to be applied before probability was. It first appeared, apparently being based on intuitive and subjective chances and in everyday life rather than in science. Maimonides (Rabinovitch 1973, p. 164): *A marriage settlement* [insurance for a woman against divorce or death of husband] *of 1000 zuz can be sold at a present value of 100*, [but] *if the face value were 100 it could not be sold for 10 but rather for less*. Large (though not more likely) gains had been considered preferable, and the same subjective tendency is existing nowadays (and the organizers of lotteries mercilessly take advantage of it). Similar ideas not quite definite either and again connected with insurance appeared in Europe a few centuries later (Sheynin 1977, pp. 206 – 209).

The theory of probability which *officially* originated in 1654, in the correspondence of Pascal and Fermat, effectively applied expectation. Here is one of their main problems which they solved independently from each other. Gamblers A and B agree to play until one of them scores 5 points (not necessarily in succession) and takes both stakes. For some reason the game is interrupted when the score was 4:3 to A. So how should they share the stakes?

Even then that problem was venerable; there are indications that a certain mathematician had solved it at least in a particular case. Note that sharing the stakes proportionally to 4:3 would have been fair when playing chess, say, i. e., when the gamblers' skill is decisive. In games of chance, however, everything depends on chance and the past cannot influence the future (cf. § 1.2.2).

Here is the solution. Gambler A has probability $p_1 = 1/2$ (Pascal and Fermat kept to chances) of winning the next play; he can also lose it with the same probability but then the score will equalize and the stakes should be equally shared. A's share (the expectation of his gain) will therefore be 1/2 + 1/4 = 3/4 of both stakes. The expectation of the second gambler is therefore 1/4 of both stakes and it could have been calculated independently.

It was a man about town, De Méré, who turned Pascal's attention to games of chance (Pascal 1654/1998, end of Letter dated 29 July 1654). He was unable to understand why the probability of an appearance of a six in 4 casts of a die was not equal to that of the appearance of two sixes in 24 casts of two dice as it followed from an old approximate rule. Here, however, are those probabilities:

$P_1 = 1 - (5/6)^4 = 0.518$, $P_2 = 1 - (35/36)^{24} = 0.492$.

So De Méré knew that gamblers had noted a difference of probabilities equal to 0.026. Cf. a similar remark made by Galileo (§ 1.1.1).

Huygens (1657) published a treatise on calculations in games of chance. He formally introduced the expectation in order to justify the sharing of stakes and the solutions of other problems. He substantiated the expediency of applying it for estimating a random variable (a random winning) by reasonable considerations.

Jakob Bernoulli (1713, part 1) however suggested a much simpler justification. Here is a quotation from Huygens and Bernoulli's reasoning (his part 1 was a reprint of Huygens complete with important comments).

Huygens, Proposition 3. Having *p* chances to get *a* and *q* chances to get *b* and supposing that all these chances are the same, I obtain

$$\frac{pa+qb}{p+q}. \qquad (2.10)$$

Since *p* and *q* are chances rather than probabilities, their sum is not unity as it was in formula (2.7). And here is Bernoulli. Suppose there are $(p + q)$ gamblers, and each of *p* boxes contains sum *a*, and each of *q* boxes contains *b*. Each gambler takes a box and all together get $(pa + qb)$. However, they are on the same footing, should receive the same sum, i. e., (2.10).

As stated in § 1.2.1, a mathematical theory cannot be based on boxes or gamblers, and even De Moivre introduced expectation axiomatically, without justifying it. And so it is being introduced nowadays, although Laplace (1814/1886, p. XVIII) just stated that it is *la seule equitable.*

Several centuries of applications have confirmed the significance of the expectation although in 1713 Nikolaus Bernoulli, in a letter to Montmort published by the latter (Montmort 1708/1713, p. 402) devised a game of chance in which it did not help at all.

Gambler A casts a die … However, the die was very soon replaced by a coin. And so, if heads appears at once, B pays A 1 écu; if heads only appears at the second toss, he pays 2 écus, 4 écus if only at the third toss etc. Required is the sum which B ought to receive beforehand.

Now, A gets 1 écu with probability 1/2, 2 écus with probability 1/4, 4 écus with probability 1/8 etc and the expectation of his gain is

$$1\cdot(1/2) + 2\cdot(1/4) + 4\cdot(1/8) + \ldots = (1/2) + (1/2) + (1/2) + \ldots = \infty. \quad (2.11)$$

However, no reasonable man will agree to pay B any considerable sum and hope for a large (much less, an infinite) gain. He will rather decide that heads will first occur not later than at the sixth or seventh toss and that he ought to pay beforehand those 1/2 écus not more than six or seven times; all the rest infinite terms of the series (2.11) will therefore disappear.

Buffon (1777, § 18) reported that 2048 such games resulted in A's mean gain of 4.9 écus and that only in 6 cases they consisted of 9 tosses, of the largest number of them. His was the first statistical study of games of chance. On a much greater scale Dutka (1988) conducted a similar investigation by applying a computer.

This paradox continued to interest mathematicians up to our time, but it was Condorcet (1784, p. 714) who left the most interesting remark: one game, even if infinite, is still only one trial; many games are needed for stochastically considering them. Freudenthal (1951) independently repeated this remark and additionally suggested that before each game the gamblers ought to decide by lot who will pay whom beforehand.

A similar statement about neglecting low probabilities holds for any game of chance (and any circumstance in everyday life). If there are very large gains in a lottery available with an extremely low probability (which the organizers will definitely ensure), they ought to be simply forgotten, neglected just like the infinite tail of the series (2.11).

But then, how low should a neglected probability be? Buffon (1777, § 8), issuing from his mortality table, suggested the value 1/10,000, the probability of a healthy man 56 years old dying within the next 24 hours. What does it mean for the Petersburg game? We have

$$1/2^n = 1/10{,}000,\ 2^n = 10{,}000,\ n\lg 2 = 4 \text{ and } n \approx 13.3.$$

Even that is too large: recall Buffon's experiment in which the maximal number of tosses only amounted to 9. This result also means that 1/10,000 was too low; we may often neglect much higher probabilities and, anyway, a single value for a neglected probability valid in any circumstances should not be assigned at all. And some events (the Earth's collision with a large asteroid) should be predicted with a probability much higher than (1 – 1/10,000). It is not however, clear how to prevent such global catastrophes.

Reader! Do you think about such probabilities when crossing the road?

While attempting to solve the paradox of the invented game, Daniel Bernoulli (1738) introduced *moral expectation* (but not the term itself). He published his memoir in Petersburg, and thus appeared the name *Petersburg game.* In essence, he thought that the real value of a gambler's gain is the less the greater is his fortune. He applied his novelty to other risky operations and for some decades it had been widely appraised (but not implemented in practice). At the end of the 19th century economists had developed the theory of marginal utility by issuing from moral expectation.

2.3.1-1. *The properties of the expectation.* **1)** Suppose that $\xi = c$ is constant. Then

$$\mathrm{E}c = \int_a^b c\varphi(x)dx = c\int_a^b \varphi(x)dx = c.$$

The expectation of a constant is that very constant.

2) The expectation of a random variable $a\xi$ is

$$\mathrm{E}a\xi = a\int_a^b x\varphi(x)dx = a\mathrm{E}\xi.$$

When multiplying a random variable by a constant its expectation is multiplied by that very constant.

3) Two (or more) random variables ξ and η are given; their densities are φ(*x*) and η(*y*) and the expectation of their sum is sought. It is equal to the double integral

$$\mathrm{E}(\xi + \eta) = \int_a^b\int_c^d (x+y)\varphi(x)\psi(y)dxdy =$$
$$\int_a^b\int_c^d x\varphi(x)\psi(y)dydx + \int_a^b\int_c^d y\varphi(x)\psi(y)dxdy.$$

Here, *c* and *d* are the extreme points of the domain of the second function and *a* and *b* have a similar meaning (see above). Notation η(*y*) instead of η(*x*) does not in essence change anything but transformations become clearer.

The first integral can be represented as

$$\int_c^d \psi(y)dy\int_a^b x\varphi(x)dx = \mathrm{E}\xi,$$

since the integral with respect to *y* is unity. Just the same, the second integral is Eη and therefore

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

The expectation of a sum of random variables is equal to the sum of the expectations of the terms. A similar statement can be proved about the difference of random variables: *its expectation is equal to the difference of the expectations of the terms*.

Note however that differences in such theorems (not only in the theory of probability) are usually not mentioned since by definition subtraction means addition of contrary magnitudes; thus, $a - c \equiv a + (-c)$.

4) Without proof: *the expectation of a product of two independent random variables equals the product of their expectations*:

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

This property is immediately generalized on a larger number of random variables.

All the properties mentioned above also take place for expectations of discrete random variables.

2.3.2. *Variance* is the second main notion characterizing distributions of random variables, their scattering. An inscription on a Soviet matchbox stated: *approximately 50 matches*. But suppose that actually one such box contains 30 matches, another one, 70. The mean is indeed 50, but is not the scattering too great? And what does *approximately* really mean?

Suppose that only some values $x_1, x_2, \ldots, x_n$ of a random variable ξ (a sample of size n) are/is known. Then the sample variance of ξ is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}. \qquad (2.12)$$

It is also called *empirical* since the values of x_i are the results of some experiment or trial.

Why function (2.12) is chosen as a measure of scattering, and why its denominator is $(n - 1)$ rather than n? I attempt at explaining it, but first I add that the variance (not sample variance) of the same ξ, varξ, of a discrete or continuous variables is, respectively,

$$\sum_{i=1}^{n} p_i(x_i - \mathrm{E}\xi)^2, \; \sigma_\xi^{\,2} = \int_a^b (x - \mathrm{E}\xi)^2 \varphi(x)dx, \qquad (2.13)$$

where a, b and $\varphi(x)$ have the same meaning as in formula (2.9).

It was Gauss (1823) who introduced the variance as a measure of the scattering of observations. Its choice, as he indicated, is more or less arbitrary, but such a measure should be especially sensitive to large errors, i. e. should include $(x - \mathrm{E}\xi)$ raised to some natural power (2, 3, …), and remain positive which excludes odd powers. Finally, that measure should be as simple as possible which means the choice of the second power of that binomial. Actually, Gauss (1823, §§ 37 – 38) had to determine only the sample variance and to apply the arithmetic mean instead of the expectation. Below, I will say more about the advantages of the variance.

Suppose that x_i, $i = 1, 2, \ldots, n$, are the errors of observation, then the sample variance will be

$$[xx]/n$$

where $[xx]$ is Gauss' notation denoting the sum of the squares of the x_i. These errors are however unknown, and we have to replace them by the deviations of the observations from their arithmetic mean. Accordingly, as Gauss proved in the sections mentioned above, the sample variance ought to be represented by formula (2.12). He (1821 – 1823/1887, p. 199) remarked that that change was also demanded by the *dignity of science*.

But suppose that a series of observations is corrupted by approximately the same systematic error. Then those formulas will not take it into considerations, will therefore greatly corrupt reality: the scattering will not perhaps be large although the observations deviated from the measured constant. Gauss himself had directly participated in geodetic observations, therefore did not trust his own formulas (because of the unavoidable systematic errors) and measured each angle until becoming satisfied that further work was useless. Extracts from his field records are published in vol. 9 of his *Werke*, pp. 278 – 281.

Not only the sample variance, but a square root of it (not only s^2, but s) is applied as well. That s is called *standard deviation*, or, in the theory of errors, *mean square error*.

And now we can specify statements similar to *approximately 50 matches in a box*. Carry out a thankless task: count the matches x_1, x_2, …, x_{10} in 10 boxes, calculate their mean $\overline{x}$ (their sample mean, since the number of such boxes is immense), the deviations $(x_1 - \overline{x})$, $(x_2 - \overline{x})$, …, $(x_{10} - \overline{x})$, and finally the sample variance (or standard deviation). A deviation of some x_i from the approximately promised value that exceeds two mean square errors is already serious.

The expectation of a random variable can be infinite, as in the case of the Petersburg game, and the same can happen with the variance. *Example.* A continuous random variable distributed according to the Cauchy law

$$\varphi(x) = \frac{2}{\pi(1+x^2)},\ 0 \le x < \infty. \qquad (2.14)$$

Note that equalities such as $x = \infty$ should be avoided since infinity is not a number but a variable. Also bear in mind that the distribution (2.14) first occurred in Poisson (1824, p. 278).

Now, the variance. It is here

$$\text{var}\xi = \frac{2}{\pi}\int_0^\infty x^2 \cdot \frac{1}{1+x^2}\,dx = \frac{2}{\pi}\left[\int_0^\infty 1 \cdot dx - \int_0^\infty \frac{1}{1+x^2}\,dx\right]$$

The second integral is

$$\text{arctg}x\Big]_0^\infty = \pi/2,$$

but the first does not exist (and the variance is infinite):

$$\int_0^\infty dx = x\Big]_0^\infty \to \infty.$$

The arithmetic mean of observations, if they are so unsatisfactory that their errors obey the Cauchy distribution, is not better than an isolated observation. Indeed, according to formula (2.16) from § 2.3.2-2 the variance of the mean of n observations is n times less than the variance of a single observation, that is, n times less than infinity and is therefore also infinite.

2.3.2-1. *A second definition of variance.* Definition (2.13b) can be written as

$$\text{var}\xi = \int_a^b x^2\varphi(x)dx - 2\int_a^b x\text{E}\xi\varphi(x)dx + \int_a^b (\text{E}\xi)^2\varphi(x)dx.$$

Now, Eξ is constant and can be separated:

$$\text{var}\xi = \int_a^b x^2\varphi(x)dx - 2\text{E}\xi\int_a^b x\varphi(x)dx + (\text{E}\xi)^2\int_a^b \varphi(x)dx.$$

By definition, the first integral is $\text{E}\xi^2$, and the second, Eξ. The third integral is unity according to the property of the density. Therefore,

$$\text{var}\xi = \text{E}\xi^2 - 2(\text{E}\xi)^2 + (\text{E}\xi)^2 = \text{E}\xi^2 - (\text{E}\xi)^2. \quad (2.15)$$

This formula is usually assumed as the main definition of variance.

2.3.2-2. *The properties of density.*

1) *The density of a sum of independent random variables*. By the second definition of variance we have

$$\text{var}(\xi + \eta) = \text{E}(\xi + \eta)^2 - [\text{E}(\xi + \eta)]^2 =$$
$$\text{E}\xi^2 + 2\text{E}(\xi\eta) + \text{E}\eta^2 - [(\text{E}\xi)^2 + 2\text{E}\xi\text{E}\eta + \text{E}\eta^2].$$

Then, according to the fourth property of expectation of independent random variables (§ 2.3.1-1),

$$\text{E}(\xi\eta) = \text{E}\xi\cdot\text{E}\eta$$

so that

$$\text{var}(\xi + \eta) = [\text{E}\xi^2 - (\text{E}\xi)^2] + [\text{E}\eta^2 - (\text{E}\eta)^2] = \text{var}\xi + \text{var}\eta.$$

The variance of a sum of independent random variables is equal to the sum of their variances.

2) *Corollary*: Variance of the arithmetic mean. Given observations $x_1, x_2, \ldots, x_n$ and their arithmetic mean (2.8b)

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

is calculated. Formula (2.12) provides the sample variance of observation x_i, but now we need the variance of the mean. By the theorems on the variance of the sum of random variables (the results of observation are random!) and on the product of a random variable by a constant (here, it is $1/n$), we obtain at once a simple but important formula

$$\text{var}\,\overline{x} = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+...+(x_n-\overline{x})^2}{n(n-1)}. \qquad (2.16)$$

The variance of the arithmetic mean of n observations is n times less than the variance of each of them.

Here, like in formula (2.12), we certainly assume that the observations are possible values of one and the same random variable.

3) *The variance of a linear function of a random variable.* Suppose that $\eta = a + b\xi$ is a linear function of random variable ξ (and therefore random as well just like any function depending on a random variable). The variance of ξ, varξ, is known and required is varη. Such problems occur often enough.

By formula (2.15)

$$\text{var}\eta = \text{E}\eta^2 - (\text{E}\eta)^2 = \text{E}(a + b\xi)^2 - [\text{E}(a + b\xi)]^2.$$

The first term is

$$\text{E}(a^2 + 2ab\xi + b^2\xi^2) = a^2 + 2ab\text{E}\xi + b^2\text{E}\xi^2.$$

The second term is

$$[\text{E}a + \text{E}(b\xi)]^2 = (\text{E}a)^2 + 2\text{E}a\text{E}(b\xi) + (\text{E}b\xi)^2 = a^2 + 2ab\text{E}\xi + b^2(\text{E}\xi)^2$$

and their difference is $b^2[\text{E}\xi^2 - (\text{E}\xi)^2]$.

According to formula (2.15) var$\eta = b^2$var ξ.

And so, an addition of a constant to a random variable does not change the variance, and, when multiplying such a variable by a constant coefficient, its variance is multiplied by the square of that constant:

$$\text{var}(a + \xi) = \text{var}\xi,\ \text{var}(b\xi) = b^2\text{var}\xi.$$

2.4. Parameters of Some Distributions

In § 2.2 we have determined the parameters of a few distributions, but the binomial and the normal laws are still left.

2.4.1. *The binomial distribution.* Suppose that μ_k is a random number of the occurrences of an event in the *k*-th trial, 0 or 1. If the probability of its happening is *p*, then

$$\text{E}\mu_k = 1\cdot p + 0\cdot q = p.$$

In a series of *n* trials that event occurs

$$(\mu_1 + \mu_2 + \ldots + \mu_n) = \mu \text{ times, E}\mu = \text{E}\mu_1 + \text{E}\mu_2 + \ldots + \text{E}\mu_n = pn.$$

Then, see formula (2.15),

$$\text{var}\mu_k = \text{E}\,\mu_k^2 - (\text{E}\mu_k)^2.$$

However, μ_k^2 takes the same values, 0 and 1, as μ_k, and with the same probabilities, p and q, so that

$$\text{var}\mu_k = p - p^2 = p(1-p) = pq,\ \text{var}\mu = \text{var}\mu_1 + \text{var}\mu_2 + \ldots + \text{var}\mu_n = pqn.$$

The magnitudes Eμ and varμ characterize the frequency μ. Recall that in § 2.2.3 we discussed the parameters of the binomial distribution proper.

2.4.2. *The normal distribution*. It follows from formula (2.4) that the form of the normal curve depends on the value of σ; the less it is, the more is the area under that curve concentrated in its central part. The values of the random variable ξ close to the abscissa of the curve's maximum become more probable, the random variable as though shrinks.

At $a = 0$ the graph of the density of the normal distribution becomes symmetrical with respect to the y-axis so that a is the *location parameter*. Note that this term is applied to any densities whose formula contains the difference $x - a$.

The analytical meaning of both parameters is very simple:

$$a = \text{E}\xi,\ \sigma^2 = \text{var}\xi. \qquad (2.17\text{a}, 2.17\text{b})$$

We will prove (2.17a) and outline the proof of (2.17b). We have

$$\text{E}\xi = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x\exp[-\frac{(x-a)^2}{2\sigma^2}]dx.$$

Now, $x = [(x - a) + a]$ and the integral can be written as

$$\frac{1}{\sigma\sqrt{2\pi}}\{\int_{-\infty}^{\infty}(x-a)\exp[-\frac{(x-a)^2}{2\sigma^2}]dx + a\int_{-\infty}^{\infty}\exp[-\frac{(x-a)^2}{2\sigma^2}dx\}.$$

In the first integral, the integrand is an odd function of $(x - a)$, which, just as x, changes unboundedly from $-\infty$ to ∞. This integral therefore disappears (the *negative* area under the x-axis located to the left of the y-axis is equal to the positive area above the x-axis located to the right of the y-axis).

Then, in the second integral, let

$$\frac{x-a}{\sigma\sqrt{2}} = z,\ dx = \sigma\sqrt{2}dz, \qquad (2.18)$$

so that it is equal to

$$\int_{-\infty}^{\infty}\exp(-z^2)dz\cdot\sigma\sqrt{2}.$$

Euler was the first to calculate it; without the multiplier σ√2 it is equal to √π. Finally, taking into account all three multipliers, a, σ√2 и $1/\sigma\sqrt{2\pi}$, we arrive at a, QED.

Now the formula (2.17b):

$$\text{var}\xi = \text{E}(\xi - \text{E}\xi)^2 = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-a)^2\exp[-\frac{(x-a)^2}{2\sigma^2}]dx.$$

We have applied here the just derived formula (2.17a). Now we ought to introduce a new variable, see (2.18), and integrate by parts.

2.5. Other Characteristics of Distributions

2.5.1. *Those replacing expectation.* For a sample (sometimes the only possibility) those characteristics replace the arithmetic mean or estimate the location of the measured constant in some other way.

2.5.1-1. *The median.* Arrange the observations x_1, x_2, …, x_n of a random variable in an ascending order and suppose that the thus ordered sequence is $x_1 \le x_2 \le \ldots \le x_n$. Its median is the middlemost observation, quite definite for odd values of n. Suppose that $n = 7$, the median will then be x_4. For even values of n the median will be the halfsum of the two middle terms; thus, for $n = 12$, the halfsum of x_6 and x_7.

For continuous random variables with density φ(x) the median is point x_0 which divides the area under the density curve into equal parts:

$$\int_a^{x_0}\varphi(x)dx = \int_{x_0}^{b}\varphi(x)dx = 1/2.$$

In other words, the median corresponds to equality $F(x) = 1/2$. To recall: the entire area under the density curve is unity; a and b are the extreme points of the domain of φ(x).

For some densities, as also when the density is unknown, the median characterizes a random variable more reliably then the arithmetic mean. The same is true if the extreme observations possibly are essentially erroneous. Indeed, they can considerably displace the mean but the median will be less influenced.

Mendeleev (1877/1949, p. 156), who was not only a chemist, but an outstanding metrologist, mistakenly thought that, when the density remained unknown, the arithmetic mean ought to be chosen.

Continuous distributions are also characterized by quantiles which correspond to some probabilities p, that is, points $x = x_p$ for which $F(x) = p$, so that the median is a quantile corresponding to $p = 1/2$. Its exact location can be not quite certain, cf. the case of the median.

2.5.1-2. *The mode.* This is the point (or these are the points) of maximal density. It (one of them) can coincide with the arithmetic mean. Accordingly, the density is called unimodal, bimodal, … or even antimodal (when a density has a point of minimum). In case of discrete random variables the mode is rarely applied.

2.5.1-3. *The semi-range* (*mid-range*). This is a very simple but unreliable measure since the extreme values can be considerably erroneous and no other observations are taken into account. It had been widely applied in the 18th century for estimating mean monthly values of meteorological elements (for example, air temperatures). It was certainly easier to calculate the mid-range than the mean of 30 values. Interestingly, Daniel Bernoulli (1778, § 10) indicated that he had found it to be *less often wrong* than [he] thought …

2.5.2. *Characteristics replacing the variance*

2.5.2-1. *The range.* The (sample) range is the difference between the maximal and the minimal measured values of a random variable, cf. § 2.5.1-3. The not necessarily equal differences ($x_n - \bar{x}$) and ($\bar{x} - x_1$) are also sometimes applied. All these differences are unreliable. In addition to the remarks in that subsection I note that they can well increase with the number of observations; there can appear a value less than x_1 or larger than x_n.

It is certainly possible to apply instead the fractions $(x_n - x_1)/n$, $(x_n - \bar{x})/n$ and $(\bar{x} - x_1)/n$. The denominator coincides with the possibly increasing number of observations but the numerator changes uncertainly. All the measures mentioned here concern a series of observations rather than a single result.

2.5.2-2. *The mean absolute error.* It, just as the probable error (see 2.5.2-3), characterizes a single observation. Denote observations by x_1, x_2, …, x_n, then the mean absolute error will be

$$\sum_{i=1}^{n} | x_i | \div n.$$

It had been applied, although not widely, when treating observations.

2.5.2-3. *The probable error.* It was formally introduced by Bessel (1816, pp. 141 – 142) as a measure of precision, but even Huygens (1669/1895), in a letter to his brother dated 28 Nov. 1669, mentioned the idea of a probable value of a random variable. Discussing the random duration of human life, he explained the difference between the expected interval (the mean value derived from data on many people) and the age *to which a person with equal probabilities can live or not.*

Both durations of life should be calculated separately for men and women, which in those times was not recognized. Women *generally* live longer and this possibly compensates them for a life more difficult both in the biological and social sense but they seem to recall this circumstance rather rarely.

Bessel had indeed applied that same idea, repeatedly found in population statistics and, for example, when investigating minor changes in the period of the swings of a pendulum (Daniel Bernoulli 1780). According to Bessel, a probable error of an observation is such that *with equally probability will be either less or larger than the really made error.*

For symmetric distributions the probable error is numerically equal to the distance between the median and the qauntile corresponding to $p = 1/4$ or $3/4$; it is the probability that an observation thus deviates in either side from the median. For the normal distribution that distance is

0.6745σ, and many authors had understood (still understand?) that relation as a universal formula or had tacitly thought that they have dealt with the normal distribution.

Moreover, I am not sure that there exists a generally accepted definition of the probable error suitable for asymmetric distributions, i. e., when the distances from the median to the quantiles $p = 1/4$ and 3/4 do not coincide. If in such cases the probable error is still meaningful, it is perhaps permissible to say that it is equal to half the distance between those quantiles.

The idea of the probable error is so natural that that measure became universally adopted whereas, perhaps until the second half of the 20th century, the mean square error had been all but forgotten. In the third (!) edition of his serious geodetic treatise Bomford (1971, pp. 610 – 611) *reluctantly* abandoned it and went over to the mean square error.

So why is the latter better? We may bear in mind that the probable error is connected with the median which is not always preferable to the arithmetic mean. Then, it, the mean square error (or, rather, the variance), is the most reliable measure. The variance (we may only discuss the sample variance) is a random variable, it therefore has its own sample variance. True, as mentioned above, a similar remark is applicable to any sample measure (in particular, to the arithmetic mean). However, unlike other measures of scattering, the variance of the variance is known, first derived by Gauss (1823, § 40). True, he made an elementary mistake corrected by Helmert (1904), then independently by Kolmogorov et al (1947).

One circumstance ought to be however indicated. Practically applied is not the variance, but its square root, the standard deviation (the mean square error); and if the variance of the variance is a, it does not at all mean that the variance of the latter is $\sqrt{a}$; for that matter, it is only known for the normal distribution, see below. Again, the sample variance is an *unbiased* estimate of the general, of the *population* variance which means that its expectation is equal to that variance, whereas the sample standard deviation has no similar property. Recall that Gauss (§ 2.3.2) remarked that the formula for the sample variance had to be changed; now I additionally state that he had thus emphasised the essential role of unbiasedness although currently it is much less positively estimated.

2.5.2-4. *An indefinite indication of scattering.* We sometimes meet with indications such as *This magnitude is equal to a* ± *c*. It can be understood as … *equal to any value between a – c* and *a + c*, but it is also possible that c is not the maximal but, for example, the probable error. And, how was that c obtained? We have approached here the important subject of interval estimation.

2.6. Interval Estimation

Denote some parameter of a function or density by λ and suppose that its sample value $\hat{\lambda}$ is obtained. Required is an estimate of the difference $|\hat{\lambda} - \lambda|$. Its *interval* estimation means that, with α and δ being indicated,

$$P(|\hat{\lambda} - \lambda| < \delta) > 1 - \alpha.$$

Now, we may state that the *confidence interval* $[\hat{\lambda}-\delta;\ \hat{\lambda}+\delta]$ covers the unknown λ with *confidence probability* (*confidence coefficient*) $(1 - \alpha)$. This method of estimation is reasonable if α is small (for example, 0.01 or 0.05, but certainly much larger than the Buffon value 1/10,000), and such that δ is also sufficiently small. Otherwise the interval estimation will show that either the number of observations was too small or that they were not sufficiently precise. Note also that in any case other observations can lead to other values of $\hat{\lambda}$ and δ.

Suppose that a constant A is determined by observations. Then, adopting simplest assumptions (Bervi 1899), we may assume that the obtained range $[x_1; x_n]$ covers it with probability

$$P(x_1 \le A \le x_n) = 1 - 1/2^{n-1}.$$

I indicated the deficiency of this trick in § 2.5.1-3. Similar conclusions were made by astronomers in the antiquity (Sheynin 1993b, § 2.1). Issuing from all the existing observations (not only his own) the astronomer selected some bounds (a and b) and stated that $a \le A \le b$. Probabilities had not been mentioned but the conclusion made was considered almost certain.

When determining a constant, any measure of scatter may be interpreted as tantamount to a confidence characteristic. Indeed, suppose that the arithmetic mean $\bar{x}$ of observations is calculated and its mean square error m determined. Then the probability $P(\bar{x} - m \le \bar{x} \le \bar{x} + m)$ can be established by statistical tables of the pertinent law of distribution as $P(0 \le \bar{x} \le \bar{x} + m) - P(0 \le \bar{x} \le \bar{x} - m)$; the difference between strict and non-strict inequalities can be neglected. So exactly that P is indeed the confidence probability and $[\bar{x} - m;\ \bar{x} + m]$, the confidence interval.

2.7. The Moments of a Random Variable

This subject can be quite properly included in § 2.6, but it deserves a separate discussion. Moments characterise the density and can sometimes establish it.

The initial moment of order s of a discrete or continuous random variable ξ is, respectively,

$$\alpha_s(\xi) = \sum_x x^s p(x),\ \ \nu_s = \int x^s \varphi(x)dx. \qquad (2.19)$$

In the first case, the summing is extended over all the values of x having probabilities $p(x)$ whereas the integral is taken within the extreme points of the domain of the known or unknown density $\varphi(x)$ of the continuous random variable.

Also applied are the central moments

$$\mu_s(\xi) = \sum_i (x_i - \mathrm{E}\xi)^s p(x_i),\ \ \mu_s(\xi) = \int (x - \mathrm{E}\xi)^s \varphi(x)dx. \qquad (2.20)$$

Both these formulas (the integral is taken between appropriate bounds) can be represented as

$$\mu_s(\xi) = E(\xi - E\xi)^s.$$

Sample (empirical) initial moments for both discrete and continuous random variables certainly coincide:

$$m_s(\xi) = \sum_i x_i^s \div n, \qquad (2.21)$$

where n is the number of measured (observed) values of ξ.

The central sample moments are

$$m_s(\xi - \overline{x}) = \frac{1}{n}\sum (x_i - \overline{x})^s.$$

The measured (observed) values are often combined within certain intervals or categories. Thus (Smirov & Dunin-Barkovski 1959/1969, § 1 in Chapter 3), 70 samples containing 5 manufactured articles each were selected for checking the size of such articles. In 55 samples the size of each of the 5 articles was standard, in 12 of them 2 were non-standard, and in 3, 1 was non-standard:

Number of samples	55	12	3
Number of defective articles	0	1	2
Frequencies of the various outcomes	0.786	0.171	0.043

Here the frequency, for example in the fist column is 55/70.

Many definitions of mathematical statistics have been offered, but only once were statistical data mentioned (Kolmogorov & Prokhorov 1974/1977, p. 721): they denote *information about the number of objects which possess certain attributes in some more or less general set.*

Those numbers above are indeed statistical data; they were separated into sets with differing numbers of defective articles in the samples. Such separation can often be made in several ways; however, if the range of the values of the random variable (the number of defective articles) is sufficiently wide (here, we have a very small range from 0 to 5, but actually even from 0 to 2), there should not be too few sets or intervals. On the other hand, there should not be too many of them either: too many subdivisions of the data is a *charlatanisme scientifique* (Quetelet 1846, p. 278)

And so, when combining the data, formula (2.21) becomes

$$m_s(\xi) = \frac{1}{n}\sum_i n_{x_i} x_i^s, \qquad (2.22)$$

where n_x is the number of the values of the random variable in interval x. In our example

$$m_s(\xi) = \frac{1}{70}(55 \cdot 0^s + 12 \cdot 1^s + 3 \cdot 2^s) = 0.786 \cdot 0^s + 0.171 \cdot 1^s + 0.043 \cdot 2^s.$$

The cases of $s = 1$ and 2 (see the very beginning of this section) coincide with expectation and variance respectively and formulas (2.20) correspond to formulas (2.13). *The first moment is the expectation, the second moment is the variance*. But is it possible and necessary to establish something about the other almost infinitely many moments? Suffice it to consider the next two of them.

Suppose that the density $\varphi(x)$ is symmetrical with respect to the y-axis. Then for odd values of s the moments

$$\nu_s = \int_{-\infty}^{\infty} x^s \varphi(x)dx = 0.$$

Indeed, in this case the integrand is the product of an odd and an even function and is therefore odd and the integral is taken between symmetrical bounds.

If some odd moment differs from zero, the density cannot be symmetric (i. e., even) and this moment will therefore characterize the deviation of $\varphi(x)$ from symmetry. But which moment should we choose as the measure of asymmetry?

All sample moments depend on the observed values of the appropriate random variable, are therefore random variables as well and possess a variance. It is also known that *the variances of the moments of higher orders are larger than those of the first few*. The moments of higher orders are therefore *unreliable*.

It is thus natural to choose the third moment as the measure of asymmetry of the density $\varphi(x)$ of a random variable; more precisely, the third sample moment since apart from observations we have nothing to go on:

$$m_3(\xi) = \frac{1}{n-1}\sum (x_i - \overline{x})^3. \tag{2.23}$$

One more circumstance. The dimensionality of the third moment is equal to the cube of the dimensionality of $(x_i - \overline{x})$. For obtaining a dimensionless measure, (2.23) should be divided by s^3, see formula (2.12). The final measure of asymmetry of $\varphi(x)$ is thus

$$s_k = m_3 \div s^3.$$

When discussing that formula (2.12), we indicated why its denominator should be $(n - 1)$ rather than n. The same cause compelled us to change the denominator in formula (2.23).

Now the fourth moment. For a normal random variable it is $3\sigma^4$, whereas the second moment is σ^2, see § 2.4.2. For that distribution we therefore have

$$\nu_4/\sigma^4 = 3.$$

If we now calculate the so-called excess (more precisely, the sample excess)

$$\varepsilon_k = m^4/s^4 - 3,$$

its deviation from 0 can be chosen as a measure of the deviation of un unknown density of distribution from the normal law (for which the excess disappears). The excess is here useful since in one or another sense the normal distribution is *usually* best. Pearson (1905, p. 181) introduced the excess when studying asymmetric laws.

In general, if the density is unknown, the knowledge of the first four moments is essential: when considering them as the corresponding theoretical moments of the density, it will be possible to imagine its type and therefore to calculate its parameters (hardly more than four of them).

To repeat: the normal law has only two parameters; therefore, if the calculated excess is sufficiently small, the unknown distribution will be determined by the first two moments. But what, indeed, is *sufficiently small*? We leave this question aside.

2.8. The Distribution of a Function of Random Variables

Suppose that random variables ξ and η have densities $\varphi_1(x)$ and $\varphi_2(y)$ and that $\eta = f(\xi)$ with a continuous and differentiable function f. The density $\varphi_1(x)$ is known and it is required to derive $\varphi_2(y)$. This problem is important and has to be solved often.

First of all, we (unnecessarily?) provide information about inverse functions and restrict our description to strictly monotone (increasing or decreasing) functions. The domain of an arbitrary function can however be separated into intervals of monotonic behaviour and each such interval can then be studied separately.

Suppose now that the function $y = f(x)$ strictly decreases on interval $[a; b]$. Turn its graph *to the left* until the y-axis is horizontal, and you will see the graph of the inverse function $x = \psi(y)$, also one-valued since the initial function was monotone. True, the positive direction of the y-axis and therefore of the argument y (yes, y, not x anymore) will be unusual. This nuisance disappears when looking with your mind's eye on the graph from the other side of the plane.

Return now to our problem. When ξ moves along $[a; b]$, the random point $(\xi; \eta)$ moves along the curve $y = f(x)$. For example, if $\xi = x_0$, then $\eta = f(x_0) = y_0$. It is seen that the distribution function (not the density) $F(y)$ of η, or $P(\eta < y)$, is

$$P(\eta < y) = P(x < \xi < b) = \int_x^b \varphi_1(x)dx = \int_x^b \varphi_1(z)dz,$$

where $(x; y)$ is a current point on the curve $y = f(x)$.

Pursuing a methodical aim, we have changed the variable in the integral above but certainly did not alter the lower bound. However, it can be expressed as a function of *y*: $x = \psi(y)$. So now

$$F(y) = P(\eta < y) = \int_{\psi(y)}^{b} \varphi_1(z)dz.$$

Differentiate both parts of this equality with respect to *y*, and obtain thus the density

$$F'(y) = \varphi_2(y) = -\varphi_1[\psi(y)]\cdot\psi'(y).$$

For a strictly *increasing* function *f*(*x*) the reasoning is the same although now it is the upper variable bound rather than the lower and the *minus* sign will disappear. Both cases can be written as

$$\varphi_2(y) = \varphi_1[\psi(y)]\cdot|\psi'(y)|.$$

Example (Ventzel 1969, p. 265).

$$\eta = 1 - \xi^3,\ \varphi_1(x) = \frac{1}{\pi(1+x^2)},\ -\infty < x < \infty.$$

Here, $\varphi_1(x)$, is the Cauchy distribution (mentioned in § 2.3.2 in a slightly different form). We have

$$x = \psi(y) = \sqrt[3]{1-y},$$

$$\psi'(y) = -\frac{1}{3\sqrt[3]{(1-y)^2}},\ \varphi_1[\psi(y)] = f[\sqrt[3]{1-y}] = \frac{1}{\pi[1+\sqrt[3]{(1-y)^2}]},$$

$$\varphi_2(y) = \frac{1}{\pi[1+\sqrt[3]{(1-y)^2}]}\ \frac{1}{3\sqrt[3]{(1-y)^2}}.$$

Such a simple function … The *y* can certainly be replaced by *x*.

2.9. The Bienaymé – Chebyshev Inequality

This is

$$P(|\xi - E\xi| < \beta) > 1 - \sigma^2/\beta^2,\ \beta > 0 \qquad (2.24)$$

or, which is evident,

$$P(|\xi - E\xi| \geq \beta) < \sigma^2/\beta^2.$$

Inequality (2.24), and therefore its second form as well, take place for any random variable having an expectation and a variance and are therefore extremely interesting from a theoretical point of view. However, exactly this property means that the inequality is rather rough (I discuss any one of them). In a way, it combines the two magnitudes, σ and β, without needing any other information.

Bienaymé (1853) established that inequality, but, unlike Chebyshev (1867 and later), did not pay special attention to his discovery since the subject of his memoir was not directly connected with it.

William Herschel (1817/1912, p. 579)

presumed that any star promiscuously chosen […] *out of* [more than 14 thousand] *is not likely to differ much from a certain mean size of them all.*

Stars unimaginably differ one from another and do not belong to a single population at all. The variance of their sizes is practically infinite, the notion of their mean size meaningless, and the inequality (2.24) cannot be applied. From another point of view, we may add: no positive data – no conclusion (*Ex nihilo nihil fit*!).

The English physician J. Y. Simpson (1847 – 1848/1871, p. 102) had similar thoughts: *The data* [about mortality after amputations] *have been objected to on the ground that they are collected from too many different hospitals and too many sources.* But […] *I believe* […] *that this very circumstance renders them more, instead of less, trustworthy.*

Chapter 3. Systems of Random Variables. Correlation

3.1. Correlation

In the first approximation it may be stated that the variable *y* is a function of argument *x* on some interval or the entire number axis if, on that interval (on the entire axis), one and only one value of *y* corresponds to each value of *x*. Such dependence can exist between random variables. For example, Bessel (1838, §§ 1 – 2): the error of a certain type of measurements is $\eta = a\xi^2$.

Less tight connections between random variables are also possible (the stature of children depending on the stature of parents). Their study is the aim of an important chapter of mathematical statistics, of the theory of correlation. That word means *comparison*. More precisely, correlation considers the change *in the law of distribution* of a random variable depending on the change of another (or other) random variable(s) and as a rule on accompanying circumstances as well.

Lacking that specification and certainly without quantitative studies of phenomena (*qualitative*) correlation had been known in antiquity. (As stated in § 1.1.3, the entire ancient science had been qualitative.) Thus, Hippocrates (1952, No. 44): *Persons who are naturally very fat are apt to die earlier than those who are slender.* Climatic belts were isolated in antiquity, but only Humboldt (1817, p. 466) connected them with mean yearly air temperatures.

Seidel (1865 – 1866), a German astronomer and mathematician, first quantitatively investigated correlation. He studied the dependence of the monthly cases of typhoid fever on the level of subsoil water, and then both on that level and the rainfall.

Galton (1889) had begun to develop the theory of correlation proper, and somewhat later Pearson followed suit. Nevertheless, it had been sufficiently improved much later. Markov (1916/1951, p. 533) disparagingly but not altogether justly declared that the correlation theory's

positive side is not significant enough and consists in a simple usage of the method of least squares for discovering linear dependences. However, not being satisfied with approximately determining various coefficients, the theory also indicates their probable errors, and enters here the realm of imagination […].

Discovering dependences, even if only linear, is important and estimation of precision is certainly necessary. Linnik (Markov 1951, p. 670) noted that in those times correlation theory had still being developed so that Markov's criticism made sense. However, Hald (1998, p. 677), without mentioning either Markov or Linnik, described Fisher's pertinent contribution of 1915 (which Markov certainly did not see) and thus refuted Linnik. Anyway, here is Slutsky's reasonable general comment (letter to Markov of 1912, see Sheynin 1999b, p. 132):

The shortcomings of Pearson's exposition are temporary and of the same kind as the known shortcomings of mathematics in the 17th and 18th centuries.

Now we shall discuss the correlation coefficient. Two random variables, ξ are η, are given. Calculate the moment

$$\mu_{\xi\eta} = \mathrm{E}[(\xi - \mathrm{E}\xi)(\eta - \mathrm{E}\eta)] = \mathrm{E}(\xi\eta) - 2\mathrm{E}\xi\ \mathrm{E}\eta + \mathrm{E}\xi\ \mathrm{E}\eta = \mathrm{E}(\xi\eta) - \mathrm{E}\xi\ \mathrm{E}\eta$$

and divide it by the standard deviations σ_ξ and σ_η to obtain a dimensionless measure, the correlation coefficient

$$r_{\xi\eta} = \frac{\mu_{\xi\eta}}{\sigma_\xi \sigma_\eta}.$$

For independent ξ and η both $\mu_{\xi\eta}$ and (therefore) $r_{\xi\eta}$ disappear. The inverse statement is not true! Cf.: a sparrow is a bird, but a bird is not always a sparrow. One case is sufficient for refuting the inverse statement here also. So suppose that the density of ξ is an even function, then $\mathrm{E}\xi = 0$ and $\mathrm{E}\xi^3 = 0$. Introduce now $\eta = \xi^2$, then $\mu_{\xi\eta} = \mathrm{E}\xi^3 - 0 = 0$, QED. It follows that (even a functional) dependence can exist between random variables when the correlation coefficient is zero.

That coefficient takes values from – 1 до 1. Correlation can therefore be negative. Example: the correlation (the dependence) between the stature and the weight of a person is positive, but between the distance from a lamp and its brightness is negative. Accordingly, we say that the correlation is direct or inverse.

3.2. The Distribution of Systems of Random Variables

Consider the probability $P(\xi < x, \eta < y)$. Geometrically, these inequalities correspond to an infinite region $-\infty < \xi < x$, $-\infty < \eta < y$, whereas analytically P is expressed by the distribution function

$$F(x; y) = P(\xi < x, \eta < y).$$

Taken separately, the random variables ξ and η have distribution functions $F_1(x)$ and $F_2(y)$ and densities $f_1(x)$ and $f_2(y)$. Thus, for an infinitely large x the inequality $\xi < x$ is identical and

$$F(+\infty; y) = F_2(y).$$

The density is also introduced here similar to the one-dimensional case:

$$P[(\xi;\eta) \text{ belongs to region } D] = \iint_D f(x;y)dxdy.$$

Function $f(x; y)$ is indeed the density. For independent ξ and η we have

$$f(x; y) = f_1(x) f_2(y).$$

For the bivariate (two-dimensional) normal law the density $f(x; y)$ is

$$\frac{1}{2\pi\sigma_x\sigma_y}\exp\{-\frac{1}{2(1-r^2)}[\frac{(x-\text{E}\xi)^2}{\sigma_\xi^2}-\frac{2r(x-\text{E}\xi)(y-\text{E}\eta)}{\sigma_\xi\sigma_\eta}+\frac{(y-\text{E}\eta)^2}{\sigma_\eta^2}]\}.$$

Apart from the previous notation, r is the correlation coefficient for ξ and η.

3.2.1. *The distribution of a sum of random variables.* Given, random variables ξ and η with densities $\varphi_1(x)$ and $\varphi_2(y)$. Required is the law of distribution of their sum $\omega = \xi + \eta$. For the distribution function of their system we have

$$F(x, y) = \int\int \varphi(x,\ y)dxdy.$$

In case of infinite domains of both functions the integration is over an infinite half-plane

$$F(x; y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\omega-x} \varphi(x; y)dy.$$

Differentiating it with respect to ω, we will have

$$F'_\omega(x; y) = \int_{-\infty}^{\infty} \varphi(x; \omega - x)dx = f(\omega),$$

or, after changing the places of x and y,

$$F'_\omega(x; y) = \int_{-\infty}^{\infty} \varphi(\omega - y; y)dy = f(\omega).$$

In case of independent random variables the distribution sought is called *composition* (of their densities). The formulas above lead to

$$f(\omega) = \int_{-\infty}^{\infty} \varphi_1(x)\varphi_2(\omega - x)dx = \int_{-\infty}^{\infty} \varphi_1(\omega - y)\varphi_2(y)dy .$$

Calculations are sometimes essentially simplified by geometrical considerations (Ventzel 1969, §§ 12.5 – 12.6). We only remark that the encounter problem (§ 1.1.2) can also be interpreted by means of the notions of random variable and density of distribution. Indeed, a random point (ξ; η) should be situated in a square with opposite vertices O(0, 0) and C(60, 60), and the sum ($\xi + \eta$) should be between two parallel lines, $y = x \pm 20$ (I have chosen 60 and 20 in that section). The distribution of that point could have ensured the derivation of the probability of the encounter. To recall, the moments ξ and η of the arrival of the two friends were independent.

Chapter 4. Limit Theorems

4.1. The Laws of Large Numbers

The statistical probability $\hat{p}$ of the occurrence of an event can be determined by the results of independent trials, see formula (1.7), whereas its theoretical probability p is given by formula (1.1).

We (§ 1.1.1) listed the shortcomings of that latter formula and only repeat now that it is rarely applicable since equally probable cases are often lacking. Consequently, we have to turn to statistical probability.

4.1.1. *Jakob Bernoulli.* Bernoulli (1713/2005, pp. 29 – 30) reasonably remarked that

Even the most stupid person […] *feels sure that the more* […] *observations are taken, the less is the danger of straying from the goal.* Nevertheless, he (p. 30) continued:

It remains to investigate whether, when the number of observations increases, […] *the probability of obtaining* [the theoretical probability] *continually augments so that it finally exceeds any given degree of certitude. Or* [to the contrary …] *that there exists such a degree of certainty which can never be exceeded no matter how the observations be multiplied.*

In other words, will the difference $|p - \hat{p}|$ continually decrease or not so that the statistical probability will not be a sufficiently good estimate of p.

Bernoulli proved that, in his restricted pattern, induction (trials) is (are) not worse than deduction: as $n \to \infty$ the difference $|p - \hat{p}|$ tends to disappear. His investigation opened up a new vast field. Nevertheless, not that difference itself, but its probability tends to zero. As $n \to \infty$

$$\lim P(|p - \hat{p}| < \varepsilon) = 1 \qquad (4.1)$$

with an arbitrarily small ε. This limit is exactly unity, not some lesser number, and induction is indeed not worse than deduction. But, wherever probabilities are involved, a fly appears in the ointment. The deviation of the statistical probability $\hat{p}$ from its theoretical counterpart can be considerable, even if rarely. A doubting Thomas can recall the example in § 1.2.3: the occurrence of an event with zero probability. Here, such an event is $|p - \hat{p}| \geq \varepsilon$. The limit of probability essentially differs from the *usual* limit applied in other branches of mathematics. Nothing similar can happen there!

Formula (4.1) is called the (weak) law of large numbers (a term due to Poisson). There also exists the so-called strong law of large numbers which removes the described pitfall, but I do not discuss it.

Strictly speaking, Bernoulli wrote out formula (4.1) in a somewhat different way, then continued his investigation. He proved that the inequality

$$P(|p - \hat{p}| \leq \varepsilon) \geq 1 - \delta, \delta > 0$$

will hold at given ε and δ as soon as *n* exceeds some *N*, which depends on those two numbers. He managed to determine how exactly *N* must increase with the tightening of the initial conditions.

His investigation was not really successful: the demanded values of *N* had later been considerably decreased (Pearson 1924; Markov 1924, p. 46 ff) mostly because it became possible to apply the Stirling formula unknown to Bernoulli. True, neither did De Moivre (§ 4.2) know that formula, but he derived it (even a bit before Stirling).

Mentioning Bernoulli's crude estimates Pearson (1925) inadmissibly compared his law with the wrong Ptolemaic system of the world. He missed its great general importance and, in particular, paid no attention to Bernoulli's existence theorem, of the very existence of the limit (4.1). It seems that Pearson did not set great store by such theorems.

In 1703 – 1705, before Bernoulli's posthumous *Ars Conjectandi* appeared, Bernoulli had exchanged letters with Leibniz; the Latin text of their correspondence are partially translated into German (Gini 1946; Kohli 1975); Bernoulli himself, without naming Leibniz, answered his criticisms in Chapter 4 of pt. 4 of his book. Leibniz did not believe that observations can ensure practical certainty and declared that the study of all the pertinent circumstances was more important than delicate calculations. Much later Mill (1843/1886, p. 353) supported this point of view:

A very slight improvement of the data by better observations or by taking into fuller considerations the special circumstances of the case is of more use, than the most elaborate application of the calculus of probabilities founded on the [previous] *data*.

He maintained that the neglect of that idea in applications to jurisprudence made the calculus of probability *the real opprobrium of mathematics*. Anyway, *considerations of the circumstances* and calculations do not exclude each other.

In a letter of 1714 to one of his correspondents Leibniz (Kohli 1975, p. 512) softened his doubts about the application of the statistical probability and mistakenly added that the late Bernoulli had *cultivated* [the theory of probability] in accordance with his, Leibniz', *exhortations*.

4.1.2. *Poisson*. Here is his qualitative definition of the law of large numbers (1837, p. 7):

Les choses de toutes natures sont soumises à une loi universelle qu'on peut appeler la loi des grands nombres. Elle consiste en ce que, si l'on observe des nombres très considérables d'événements d'une même nature, dépendants de causes constantes et de causes qui varient irrégulièrement, tantôt dans un sens, tantôt dans l'autre, c'est-à-dire sans que leur variation soit progressive dans aucun sens déterminé, on trouvera, entre ces nombres, des rapports à très peu près constantes.

This is a diffuse definition of a principle rather than law. And here is a contemporary qualitative definition of that law (Gnedenko 1954, § 30, p. 185): it is

The entire totality of propositions stating with probability, arbitrarily close to unity, that there will occur some event depending on an indefinitely increasing number of random events each only slightly influencing it.

The equality

$$\lim P(|\frac{\mu}{n} - \overline{p}| < \varepsilon) = 1, n \to \infty \qquad (4.2)$$

is now called the Poisson theorem. Here, μ/n is the frequency of an event in n independent trials and p_k (from which $\overline{x}$ is calculated) is the probability of its occurrence in trial k. Note that for the *Bernoulli trials* the probability of the occurrence of the studied event was constant (not p_k but simply p). Unlike formula (4.1), the new equality is general and therefore much more applicable.

4.1.3. *Subsequent history*. Chebyshev (1867) proved a more general theorem and Khinchin (1927) managed to generalize it still more. Finally, I provide another, not quite general formula for the law of large numbers: if

$$\lim P(|\overline{\xi}_n - a| < \varepsilon) = 1, n \to \infty,$$

where a is some number, the sequence of magnitudes ξ_k obeys that law.

4.2. The De Moivre – Laplace Theorem

Suppose that a studied event occurs in each trial with probability p and does not occur with probability q, $p + q = 1$ and that in n such independent trials it happened μ times. Then, as $n \to \infty$,

$$\lim P(a \le \frac{\mu - np}{\sqrt{npq}} \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp(-\frac{z^2}{2}) dz. \qquad (4.3)$$

In the limit, the binomial distribution thus becomes normal. This is what De Moivre proved in 1733 for the particular case of $p = q = 1/2$ (in his notation, $a = b = 1/2$), but then he correctly stated that a transition to the general case is easy; furthermore, the heading of his (privately printed Latin) note mentioned the binomial $(a + b)^n$. To recall, $np = E\mu$ and $npq = \mathrm{var}\mu$. Note also that the formula (4.3) is a particular case of the central limit theorem (§ 2.2.4).

Like other mathematicians of his time, De Moivre applied expansions into divergent series, only took into account their first terms, and neglected all the subsequent terms as soon as they began to increase (as soon as the series really began to diverge).

Laplace (1812, Chapter 3) derived the same formula (4.3) by means of a novelty, the Euler – Maclaurin summation formula. Furthermore, he added a term taking account of the inaccuracy occurring because of the unavoidable finiteness of *n*. Markov (1914/1951, p. 511), certainly somewhat mistakenly, called the *integral* after De Moivre and Laplace. That name persisted in Russian literature although, tacitly, in the correct way, as describing the integral theorem due to both those scholars. There also exists the corresponding local theorem

$$P(\mu) \approx \frac{1}{\sqrt{2\pi npq}} \exp[-\frac{(\mu - np)^2}{2npq}]. \qquad (4.4)$$

Assigning some μ in the right side of this formula and inserting the appropriate values of *n, p, q*, we will approximately calculate the probability of that μ. Exponential functions included in formulas (4.3) and (4.4) are tabulated in many textbooks.

A few additional remarks. Formula (4.3) describes a uniform convergence with respect to *a* and *b* (those interested can easily find this term), but Laplace (or certainly De Moivre) did not yet know that notion. Again, strict inequalities had not been then distinguished from non-strict ones. In formula (4.3), we should now apply a strict inequality in the second case (… < *b*). Then, the convergence to the normal law worsens with the decrease of *p* or *q* from 1/2 which is seen in a contemporary proof of the theorem (Gnedenko 1954, § 13).

In 1738 De Moivre included his own English translation of his private note in the second edition of the *Doctrine of Chances* and reprinted it in an extended form in the last edition (1756) of that book. However, the English language was not generally known by scientists on the Continent and the proof of (4.3) was difficult to understand since English mathematicians had followed Newton in avoiding the symbol of integral. Finally, Todhunter (1865, p. 192 – 193), the most eminent historian of probability of the 19th century, described the derivation of the formula (4.3) rather unsuccessfully and did not notice its importance. He even stated that De Moivre had only proved it for the particular case of $p = q = 1/2$. De Moivre's theorem only became generally known by the end of the 19th century.

Already in 1730 De Moivre independently derived the Stirling formula; the latter only provided him the value of the constant, $\sqrt{2\pi}$. Both Pearson (1924) and Markov (1924, p. 55 note) justly remarked that the Stirling formula ought to be called after both authors. I additionally remark that in 1730 De Moivre had compiled a table (with one misprint) of lg*n*! for $n = 10(10)900$ with 14 decimal points; 11 or 12 of them are correct.

Suppose now that it is required to determine the probability of casting a six 7 times in 100 rolls of a die. We have $p = 1/6$ and $n = 100$, then $np = 16.7$ and $\sqrt{npq} = 13.9$. By formula (4.4)

$$P(\mu = 7) \approx \frac{1}{\sqrt{2\pi}\sqrt{npq}}\exp[-\frac{(7-np)^2}{2npq}].$$

I have isolated the factor $\sqrt{2\pi}$ since the exponential function is tabulated together with $1/\sqrt{2\pi}$.

Another point. In § 2.2.4 I mentioned that De Moivre had studied the sex ratio at birth. Now I say that exactly this subject (rather topical as the following shows) became the immediate cause for the derivation of formula (4.3).

Arbuthnot (1712) collected the data on births (or rather on baptisms) in London during 1629 – 1710. He noted that during each of those 82 years more boys had been born than girls and declared that that fact was *not the effect of chance, but Divine Providence, working for a good end* since mortality of boys and men was higher than that of females and since the probability of the observed fact was only $(1/2)^{-82}$.

His reasoning was not good enough but the problem itself proved extremely fruitful. Baptisms were not identical with births, London was perhaps an exception, Christians possibly somehow differed from other people and the comparative mortality of the sexes was not really studied. Then, by itself, an insignificant probability had not proved anything and it would have been much more natural to explain the data by a binomial distribution.

In a letter of 1713 Nikolaus Bernoulli (Montmort 1708/1713, pp. 280 – 285) had indeed introduced that distribution. Denote the yearly number of births by n, μ of them boys, the unknown sex ratio at birth by m/f and $p = m/(m + f)$. Bernoulli indirectly derived the approximate equality (lacking in Bernoulli's letter)

$$P(\frac{|\mu - np|}{\sqrt{npq}} \le s) \approx 1 - \exp[-\frac{s^2}{2}],$$

where s had order $\sqrt{n}$, see Sheynin (1968; only in its reprint). He thus effectively arrived at the normal law much earlier than De Moivre.

Youshkevich (1986) reported that three mathematicians concluded that Bernoulli had come close to the local theorem (4.4) although I somewhat doubt it and the very fact that three mathematicians had to study Bernoulli's results testifies that these are difficult to interpret.

The initial goal of the theory of probability consisted in separating chance and design. Indeed, Arbuthnot, Nikolaus Bernoulli and De Moivre pursued this very aim. The last-mentioned devoted the first edition of his *Doctrine of Chances* to Newton and reprinted this dedication in the third edition of that book (p. 329). He attempted to work out, to *learn* from Newton's *philosophy*, *A method of calculating the effects of chance* [… and to fix] *certain rules for estimating how far some sort of events may rather be owing to design rather than chance …*

Note that De Moivre then did not yet prove his limit theorem.

4.3. The Bayes Limit Theorem

His main formula was (1764)

$$P(b \le r \le c) = \int_b^c u^p(1-u)^q dx \div \int_0^1 u^p(1-p)^q dx. \qquad (4.5)$$

Bayes derived it by applying complicated logical constructions, but I interpret its conditions thus: given, a unit interval and segment [*b*; *c*] lying within it. Owing to complete ignorance (Scholium to Proposition 9), point *r* is situated with equal probability anywhere on that interval; in $n = p + q$ trials that point occurred *p* times within [*b*; *c*] and *q* times beyond it.

In other words, Bayes derived the posterior distribution of a random variable having a uniform prior distribution. The denial of that assumption (of the uniform distribution) led to discussions about the Bayes memoir (§ 1.1.1-5). In addition, the situation of point *r* is not at all random but unknown. Thus, the formula (4.5) should not be applied for deriving the probability of a certain value of a remote digit in the expansion of π (Neyman 1938a/1967, p. 337).

At that time there did not exist any clear notion of density; now, however, we may say that the formula (4.5) does not contradict its definition. Bayes derived the denominator of the formula and thus obtained the value of the beta-function (Euler). Both the pertinent calculation and the subsequent work were complicated and not easy to retrace. However, Timerding, the editor of the German version of Bayes' memoir (1908), surmounted the difficulties involved. Moreover, he invented a clever trick and wrote out the result as a limit theorem. For large *p* and *q* he arrived at

$$\lim P(a \le \frac{\overline{p} - p/n}{\sqrt{pq/n^3}} \le b) = \frac{1}{\sqrt{2\pi}}\int_a^b \exp(-\frac{z^2}{2})\, dz,\ n \to \infty. \qquad (4.6)$$

Here $\overline{p}$ is a statistical estimate of the unknown probability *p* that point *r* is within [*b*; *c*], and $p/n = \mathrm{E}\overline{p}$, $pq/n^3 = \mathrm{var}\,\overline{p}$.

A comparison of the formulas (4.3) and (4.5) convinces us that they describe the behaviour of differing random variables

$$\frac{\xi_i - \mathrm{E}\xi_i}{\sqrt{\mathrm{var}\xi_i}},\ i = 1 \text{ (De Moivre)},\ i = 2 \text{ (Bayes)}.$$

The variance in the Bayes formula is larger. The proof is not really needed; indeed, statistical data are present in both cases, but additional information (the theoretical probability) is only given in formula (4.3). And it is extremely interesting that Bayes, who had no idea about the notion of variance, understood that the De Moivre formula did not describe good enough the determination of that theoretical probability by its statistical counterpart. Both Jakob Bernoulli and De Moivre mistakenly stated the opposite, but Price, an eminent statistician who communicated (and extended) the posthumous Bayes memoir, mentioned this circumstance.

But why did not Bayes himself represent his result as a limit theorem? In another posthumous contribution of the same year, 1784, Bayes clearly indicated, for the first time ever, that the application of divergent series (in particular, by De Moivre) is fraught with danger. Timerding, it ought to be remarked, managed to avoid them. Note however that such series are still cautiously applied.

I believe that Bayes had completed the first version of the theory of probability which included the Jakob Bernoulli law of large numbers and the theorems due to De Moivre and Bayes himself. In addition, Bayes was actually the main predecessor of Mises (which the latter never acknowledged). See also Sheynin (2010b).

Chapter 5. Random Processes. Markov Chains

5.1. Random Functions

Random functions are random variables changing discretely or continuously in time; for example, unavoidable noise occurring during the work of many instruments. Fixing some definite moment, we obtain the corresponding random variable, a *section of a random function*.

The law of distribution of a random function is naturally a function of two arguments one of which is time. For this reason the expectation of a random function is not a number but a (usual rather than a random) function. When fixing the moment of time, the expectation will pertain to the corresponding section of the random function and a similar statement concerns the variance. Another new point has to do with the addition of dependent random functions: the notion of correlation ought to be generalized.

A *random function without after-effect* is such for which there exists a definite probability of its transferring from a certain state to another one in such a way that additional information about previous situations does not change that probability. A good example is the Brownian motion (discovered by the English botanist Brown in 1827), the motion of tiny particles in a liquid under the influence of molecular forces.

About half a century ago, a new important phenomenon, the chaotic motion differing from random motion, began to be studied. However complicated and protracted is a coin toss, its outcomes do not change and neither do their probabilities. Chaotic motion, on the other hand, involves a rapid increase of its initial instability (of the unavoidable errors in the initial conditions) with time and countless positions of its possible paths.

It was Laplace (1781; 1812, § 15) who introduced subjective opinions (end of § 1.1.1-6) and, actually, a random process. Suppose that some interval is separated into equal or unequal parts and perpendiculars are erected from their ends. Let there be *i* perpendiculars, their total length unity, forming a non-increasing sequence in one of the two directions. Their ends are connected by a broken line and a proper number of curves, and all this construction is repeated *n* times after which the mean values of the current perpendiculars are calculated.

Laplace supposes that the lengths of the perpendiculars are assigned by *n* different people and that the worth of candidates or the significance of various causes can thus be ordered in a non-increasing order. Each curve can be considered a random function; their set, a random process; and the mean curve, its expectation. True, the calculations occurred very complicated.

Evolution according to Darwin provides a second example. Consider a totality of individuals of, say, the male sex, of some species. Each individual can be theoretically characterised by the size of its body and body parts; the unimaginable multitude *n* of such sizes is of no consequence. Introduce the usual definition of distance in an *n*-dimensional space and each individual will be represented by its point. The same space will contain the point or the subspace U of the sizes optimal for the chosen species. In the next generation, the offspring of any parents will be the better adapted to life the nearer they are to U which means that, in spite of the random scattering of the offspring around their midparents (a term due to Pearson), one generation after another will in general move towards U. However, that U will also move according to the changes in our surrounding world (and, if that movement is too rapid, the species can disappear). And so, individuals remote from U will in general perish or leave less offspring and our entire picture can be understood as a discrete random process with sections represented by each generation.

Our pattern is only qualitative; indeed, we do not know any numbers, any probabilities, for example, the probability of the mating of two given individuals of different sexes and we are therefore ignorant of any information about their offspring. Moreover, Darwin reasonably set great store by the habits of animals about which we are ignorant as well. Finally, there exists correlation between body parts of an individual. Darwin himself (1859/1958, p. 77) actually compared his theory (or, rather, hypothesis) with a random process:

Throw up a handful of feathers, and all fall to the ground according to definite laws; but how simple is the problem where each shall fall compared with problems in the evolution of species.

Opponents of evolution mostly cited the impossibility of its being produced by chance, by uniform randomness which once again showed that for a very long time that distribution had been considered as the only one describing randomness. Baer (1873, p. 6) and Danilevsky (1885, pt. 1, p. 194) independently mentioned the philosopher depicted in *Gulliver's Travels* (but borrowed by Swift form Raymond Lully, 13th – 14th centuries). That inventor, hoping to learn all the truths, was putting on record each sensible chain of words that appeared from among their uniformly random arrangements. Note that even such randomness does not exclude the gradual movement of the generations to U (but the time involved will perhaps be enormous).

Evolution began to be studied anew after Mendel's laws have been unearthed (after about 40 years of disregard) and, once more anew, after the important role of mutations has been understood.

5.2. Markov Chains (Processes with Discrete Time)

Suppose that one and only one of the events $A_1^{(s)}, A_2^{(s)}, ..., A_k^{(s)}$ occurs in trial *s* and that in the next trial the (conditional) probability of event $A_i^{(s+1)}$ depends on what happened in event *s*, but not on those preceding *s*. These conditions, if fulfilled in any trial, determine a *homogeneous Markov chain*.

Denote the conditional probability of $A_j^{(s+1)}$ as depending on $A_i^{(s)}$ by p_{ij}, then the process described by such chain is determined by a square matrix (table) of such probabilities, the *transition matrix*. Its first row is $p_{11}, p_{12}, \dots, p_{1k}$, the second one, $p_{21}, p_{22}, \dots, p_{2k}, \dots$, and the last one, $p_{k1}, p_{k2}, \dots, p_{kk}$, and the sum of the probabilities, *the transition probabilities*, in each row is unity.

It is possible to construct at once both a transition matrix for *n* trials and the limiting matrix which exists (that is, the corresponding limiting probabilities exist) if for some *s* all the elements of the matrix are positive. Markov derived this result and discovered some other findings which were later called ergodic theorems. In particular, it occurred that under certain conditions all the limiting probabilities are identical.

This remarkable property can explain, for example, the uniform distribution of the small planets along the ecliptic: a reference to these limiting probabilities which do not depend on the initial probabilities would have been sufficient. Actually, however, the small planets (more precisely, all planets) move along elliptical orbits and in somewhat differing planes.

Poincaré (1896/1987, p. 150), who had not referred to any Russian author, not even to Laplace or Poisson, justified this fact although in a complicated way. (Also, by introducing hypercomplex numbers, he proved that after a lot of shuffling the positions of the cards in a pack tended to become equally probable.)

Markov himself only applied his results to investigate the alternation of consonants and vowels in the Russian language (Petruszewycz 1983). He possibly obeyed his own restriction (Ondar 1977/1981, p. 59, Markov's letter to Chuprov of 1910):

I shall not go a step out of that region where my competence is beyond any doubt.

The term itself, *Markov chain*, first appeared (in French) in 1926 (Bernstein 1926, first line of § 16) and pertained to Markov's investigations of 1906 – 1913. Some related subjects are Brownian motion, extinction of families, financial speculation, random walk.

The urn problem discussed below can be understood as a (one-dimensional) random walk, as a discrete movement of a particle in one or another direction along some straight line with the probabilities *p* and *q* of movement depending on what had happened in the previous discrete moment. Diffusion is a similar but three-dimensional process, but a random walk with constant *p* and *q*, like the *walk* of the number of winnings of one of the two gamblers in a series of games, is not anymore a Markov chain.

And so, we will discuss the urn problem of Daniel Bernoulli (1770) and Laplace which is identical to the celebrated Ehrenfests' model (1907) considered as the beginning of the history of discrete random processes, or Markov chains. The first urn contains *n* white balls, the second urn, the same number of black balls. Required is the (expected) number of white balls in the first urn after *r* cyclic interchanges of a ball.

In his second problem Bernoulli generalized the first by considering three urns and balls of three colours. He managed to solve it elegantly, and discovered the limiting situation, an equal (expected) number of balls of each colour in each urn. A simplest method of confirming this result consists in a reference to the appropriate ergodic theorem for homogeneous Markov chains, but first we should prove that this Bernoulli problem fits the pattern of that theorem. It is not difficult. Indeed, for example, in the case of two urns, four events are possible at each interchange and the probability of each event changes depending on the results of the previous interchange. These four events are: white balls were extracted from each urn; a white ball was extracted from the first urn and a black ball from the second etc.

Laplace (1812, chapter 3) generalized the Bernoulli problem (but did not refer to him) by admitting an arbitrary initial composition of the urns, then (1814/1886, p. LIV) adding that *new urns are placed amongst the original urns*, again with an arbitrary distribution of the balls. He (p. LV) concluded, probably too optimistically, that

On peut étendre ces résultats à toutes les combinaisons de la nature, dans lesquelles les forces constantes dont leurs éléments sont animés établissent des modes réguliers d'action, propres à faire éclore du sein même du chaos des systèmes régis des lois admirables.

Chapter 6. The Theory of Errors and the Method of Least Squares

6.1. The Theory of Errors

This term (in German) is due to Lambert (1765, § 321). It only became generally used in the middle of the next century; neither Laplace, nor Gauss ever applied it although Bessel did. The theory of errors studies errors of observation and their treatment so that the method of least squares (MLSq) belongs to it. I have separated that method owing to its importance.

The theory of errors can be separated into a stochastic and a determinate part. The former studies random errors and their influence on the results of measurements, the action of the round-off errors and, the dependence between obtained relations. The latter investigates the patterns of measurement for a given order of errors and studies methods of excluding systematic errors (or minimizing their influence).

Denote a random error by ξ, its expectation will then be $E\xi = 0$. Otherwise (as it really is) ξ is not a purely random error and $E\xi = a$ is the systematic error. It shifts the even density of random errors either to the right (if $a > 0$), or to the left (if $a < 0$).

From 1756 (Simpson, § 1.2.3) until the 1920s the stochastic theory of errors, as stated there, had remained the most important field of application of the theory of probability. In a posthumous publication Poincaré(1921/1983, p. 343) noted that *La théorie des erreurs était naturellement mon principal but* and Lévy (1925, p. vii) strongly indicated that without the theory of errors his contribution on stable laws *n'aurait pas de raison d'être*.

Stable laws became an essential notion of the theory of probability, but for the theory of errors they are absolutely useless. As a corollary to the definition of a stable law it follows that the sum $\sum\xi_i$ and the arithmetic mean $\overline{\xi}$ have the same distribution as the independent and identically distributed random variables ξ_i, and Lévy proved that a real estimation of the precision of those functions of random variables, if their distribution is not stable, is very difficult. However, an observer can never know whether the errors of his measurements obey a stable law or not. Moreover, the Cauchy law is also stable, but does not possess any variance (§ 2.3.2).

In turn, mathematical statistics took over the principles of maximal likelihood and least variance (see § 6.3 below) from the stochastic theory of errors.

Now the determinate theory of errors. Hipparchus and Ptolemy could not have failed to be ignorant about them (in the first place, about those caused by the vertical refraction). Nevertheless, it was Daniel Bernoulli (1780) who first clearly distinguished random and systematic errors although only in a particular case.

Also in antiquity astronomers had been very successfully observing under the most favourable conditions. A good example is an observation of the planets during their *stations*, that is, during the change of their apparent direction of motion, when an error in registering some definite moment least influences the results of subsequent calculations. Indeed,

One the most admirable features of ancient astronomy [was] *that all efforts were concentrated upon reducing to a minimum the influence of the inaccuracy of individual observations with crude instruments by developing* […] *the mathematical consequences of very few elements* [of optimal circumstances] (Neugebauer 1950/1983, p. 250).

Actually, however, the determinate theory of errors originated with the differential calculus. Here is a simplest geodetic problem. Two angles, α and β, and side *a* are measured in a triangle and the order of error of these elements is known. Required is the order of error in the other (calculated) elements of the triangle, and thus the determination of the optimal form of the triangle.

Denote the length of any of the calculated sides by *W*. It is a function of the measurements:

$$W = f(a;\ \alpha;\ \beta),$$

and its differential, approximately equal to its error, is calculated by standard formulas.

From studying isolated geodetic figures the determinate theory moved to investigating chains and even nets of triangles. And here is a special problem showing the wide scope of that theory (Bessel 1839). A measuring bar several feet in length is supported at two points situated at equal distances from its middle. The bar's weight changes its length and the amount of change depends on the position of the supporting points. Where exactly should you place these points so that the bar's length will be least corrupted? Bessel solved this problem by means of appropriate differential equations. For a contemporary civil engineer such problems are usual, but Bessel was likely the first in this area.

Gauss and Bessel originated a new stage in experimental science. Indeed, Newcomb (Schmeidler 1984, pp. 32 – 33) mentioned the *German school of practical astronomy* but mistakenly only connected it with Bessel. True, the appropriate merits of Tycho Brahe are not known adequately. Newcomb continued:

Its fundamental idea was that the instrument is indicted […] *for every possible fault, and not exonerated till it has proved itself corrected in every point. The methods of determining the possible errors of an instrument were developed by Bessel with ingenuity and precision of geometric method …*

Gauss had detected the main systematic errors of geodetic measurements (those caused by lateral refraction, by the errors of graduating the limbs of the theodolites, and inherent in some methods of measurement) and outlined the means for eliminating/decreasing them.

For a more detailed description of this subject see Sheynin (1996, Chapter 9).

6.2. The True Value of a Measured Constant

Many sciences and scientific disciplines have to measure constants; metrology ought to be mentioned here in the first place. But what should we understand as a true value of a constant? Is it perhaps a philosophical term?

Fourier (1826) suggested its definition undoubtedly recognized earlier even if tacitly: the true value of a constant is the limit of the arithmetic mean of its n measurements as $n \to \infty$. It is easy to see that the Mises' frequentist definition of probability (§ 1.1.3) is akin to Fourier's proposal. Fourier also stated that the measurements ought to be carried out under identical conditions which was really essential for metrology but inadmissible in geodesy: differing (but good enough) external conditions were necessary for some compensation of systematic errors.

I failed to find a single reference to his definition but many authors repeated it independently from him or one another. One of them (Eisenhart 1963/1969, p. 31) formulated the unavoidable corollary: the mean residual systematic error had to be included in the *true value*:

The mass of a mass standard is […] *specified* […] *to be the mass of the metallic substance of the standard plus the mass of the average volume of air adsorbed upon its surface under standard conditions*.

Statisticians have done away with true values and introduced instead parameters of densities (or distribution functions) and their properties. A transition to more abstract notions is a step in the right direction (cf. end of § 1.2.3), but they still have to mention true values; Gauss (1816,

§§ 3 и 4) even discussed the true value of a measure of error, of something not existing in nature. For more detail see Sheynin (2007).

6.3. The Method of Least Squares

This standard method of treating observations is usually regarded as a chapter of mathematical statistics rather than probability.

Suppose that the unknown constants x and y are connected with observations $w_1, w_2, \ldots, w_n$ by linear equations

$$a_i x + b_i y + \ldots + w_i = 0,\ i = 1, 2, \ldots, n. \qquad (6.1)$$

In the general case the number of the unknowns, k, is arbitrary, but if $k > n$, the solution of (6.1) is impossible, and if $k = n$, no special methods of its solution are needed. Therefore, $k < n$. The coefficients a_i, b_i, … ought to be provided by the appropriate theory, and the system (6.1) can be supposed linear if the unknowns are small.

Indeed. Suppose that a system is not linear and that its first equation is

$$a_1 x^2 + b_1 y^2 + w_1 = 0.$$

We actually know the approximate value of the unknowns, x_0 and y_0, so that, for example, the first term of our equation is $a_1(x_0 + \Delta x)^2$ with an unknown small Δx. The term $(\Delta x)^2$ can be neglected and that first term will be $a_1(x_0^2 + 2x_0\Delta x)$. The unknown magnitude is now linear, $2a_1x_0\Delta x$, and the second term of our equation, b_1y^2, can be *linearized* in a similar way.

And now the main question: how to solve the system (6.1)? Observations are supposed to be independent (or almost so) and rejecting the $(n - k)$ redundant equations (which exactly?) would have been tantamount to rejecting worthy observations. A strict solution is impossible: any set $(x, y, \ldots)$ will leave some residual free terms (call them v_i). We are therefore obliged to impose one or another condition on these residuals. It became most usual to choose the condition of least squares

$$v_1^2 + v_2^2 + \ldots + v_n^2 = \min, \qquad (6.2)$$

hence the name, MLSq. And

$$v_i^2 = (a_i x + b_i y + \ldots + w_i)^2.$$

We ought to determine the values of $x, y, \ldots$, leading to condition (6.2), and these unknowns are therefore considered here as variables. We have

$$\frac{\partial v_i^2}{\partial x} = 2a_i(a_i x + b_i y + \ldots + w_i).$$

According to the standard procedure,

$$\frac{\partial v_1^2}{\partial x}+\frac{\partial v_2^2}{\partial x}+...+\frac{\partial v_n^2}{\partial x} = 2\sum(a_ia_ix + a_ib_iy + \ldots + a_iw_i) = 0,$$

so that, applying the Gauss elegant notation (§ 2.3.2),

$$[aa]x + [ab]y + \ldots + [aw] = 0.$$

Differentiating v_i^2 with respect to y, we similarly get

$$[ab]x + [bb]y + \ldots + [bw] = 0.$$

The derived equations are called *normal*, and it is clear that their number coincides with the number of the unknowns (yes, they became again unknown); the system of normal equations can therefore be solved in any usual way. Note, however, that the solution provides a certain set $\hat{x}, \hat{y},...$, a set of estimators of $\{x, y, \ldots\}$, of magnitudes which will remain unknown. Even the unknowns of the system of normal equations already are $\hat{x}, \hat{y},...$ rather than $x, y, \ldots$ It is also necessary to estimate the errors of $\hat{x}, \hat{y},...$, but we leave that problem aside.

Condition (6.2) ensures valuable properties to those estimators (Petrov 1954). It corresponds to the condition of minimal variance, to

$$m^2 = \frac{v_1^{\ 2} + v_2^{\ 2} + ... + v_n^{\ 2}}{n-k} = \min. \qquad (6.3)$$

The denominator is the number of redundant observations; the same is true for the formula (2.12) which corresponds to the case of one single unknown. For this case system (6.1) becomes simpler,

$$a_ix + w_i = 0,$$

and it is easy to verify that it leads to the generalized arithmetic mean.

Classical systems (6.1) had two unknown parameters of the ellipsoid of rotation best representing the figure of the Earth. After determining the length of one degree of a meridian in two different and observed latitudes it became possible to calculate those parameters whereas redundant *meridian arc measurements* led to equations of the type of (6.1). They served as a check of field measurements, they also heightened the precision of the final result (and to some extent compensated local irregularities of the figure of the Earth).

The lengths of such arcs in differing latitudes were certainly different and thus indicated the deviation of that figure from a circumference.

Legendre (1805, pp. 72 – 73; 1814) recommended the MLSq although only justifying it by reasonable considerations. Moreover, he (as also Laplace) mistakenly called the v_i 's errors of measurements and, finally, according to the context of his recommendation, he thought that the MLSq led to the minimal value of the maximal $|v_i|$. Actually, this condition is ensured by the method of *minimax*, see § 6.4.

Gauss had applied the MLSq from 1794 or 1795 and mistakenly thought that it had been known long ago. In general, Gauss did not hurry to publish his discoveries; he rather connected priority with the finding itself. He (1809, § 186) therefore called the MLSq *unser Princip* which deeply insulted Legendre. Note, however, that, unlike Legendre, Gauss had justified the new method (but later substantiated it in a different way since then, in 1809, the normal law became the only law of error).

As I see it, Legendre could have simply stated in some subsequent contribution that no one will agree that Gauss was the author of the MLSq. However, French mathematicians including Poisson (see below a few words about Laplace) sided with Legendre's opinion and, to their own great disadvantage, ignored Gauss' contributions on least squares and the theory of errors.

Much later Gauss (letter to Bessel of 1839; *Werke*, Bd. 8, pp. 146 – 147) explained his new attitude towards the MLSq:

Ich muß es nämlich in alle Wege für weniger wichtig halten, denjenigen Wert einer unbekannten Größe auszumitteln, dessen Wahrscheinlichkeit die größte ist, die ja doch immer unendlich klein bleibt, als vielmehr denjenigen, an welchen sich haltend man das am wenigsten nachteilige Spiel hat.

In other words, an integral measure of reliability (the variance) is preferable to the principle of maximal likelihood which he applied in 1809. Then, in 1809, Gauss did not refer either to Lambert (1760, § 295) or to Daniel Bernoulli (1778). The former was the first to recommend that principle for an indefinite density distribution. He had only graphically shown that density; it was a more or less symmetrical unimodal curve of the type $\varphi(x - \hat{x})$, where $\hat{x}$ can be understood as a location parameter. For observations $x_1, x_2, \ldots, x_n$ Lambert recommended to derive $\hat{x}$ from the condition (of maximum likelihood nowadays applied in mathematical statistics)

$$\varphi(x_1 - \hat{x})\, \varphi(x_2 - \hat{x}) \ldots \varphi(x_n - \hat{x}) = \max.$$

So Gauss assumed that the arithmetic mean of observations was at the same time the most probable (in the sense of maximum likelihood) estimator and discovered that only the normal distribution followed from this assumption.

In 1823 Gauss published his second and final justification of the MLSq by the principle of minimal variance (see above his letter to Bessel of 1839). Unlike his considerations in 1809, his reasoning which led him to equations (6.2) was very complicated whereas the law of error was indeed more or less normal. Thus, Maxwell (1860) proved (non-rigorously) that the distribution of gas velocities appropriate to a gas in equilibrium was normal; Quetelet (1853, pp. 64 – 65) maintained that the normal law governed the errors *faites par la nature*.

No wonder that Gauss' first formulation of the MLSq persisted (and perhaps is still persisting) in spite of his opinion. I (2012) noticed, however, that Gauss actually derived formula (6.3) as representing the minimal value of the (sample) variance independently from his complicated considerations and that, when taking this into account, his memoir (1823) becomes much easier to understand. And facts showing that the normal law was not universal in nature continued to multiply so that that memoir should be considered much more important.

The first serious opponent of the normal law was Newcomb (1886, p. 343) who argued that the cases of normal errors were quite exceptional. For treating long series of observations he recommended a mixture of differing normal laws, but the calculations proved complicated whereas his pattern involved subjective decisions. Later Eddington (1933, § 5) proved that that mixture was not stable.

Bessel (1818) discussed Bradley's series of 300 observations and could have well doubted the existence of an appropriate normal law. He noticed that large errors had appeared there somewhat more often than expected but somehow explained it away by the insufficient (!) number of observations. Much later he (1838) repeated his mistake. I (2000) noted many other mistakes and even absurdities in his contributions.

Unlike other French mathematicians, Laplace objectively described Legendre's complaint: he was the first to publish the MLSq, but Gauss had applied it much earlier. However, Laplace never recognized the utmost importance of Gauss' second substantiation of the method. Instead, he persisted in applying and advocating his own version of substantiating it. He proved several versions of the central limit theorem (CLT) (§ 2.2.4), certainly, non-rigorously (which was quite understandable) and very carelessly listing its conditions, then declared that the errors of observation were therefore normal. Laplace (1814/1886, p. LX) maintained that his finding was applicable in astronomy where long series of observations are made; cf., however, Newcomb's opinion above. Then he (1816/1886, p. 536) stated that the CLT holds in geodesy since, as it followed from his reasoning, the order of two main errors inherent in geodetic observations have been equalized. Here again he did not really take into account the conditions of that theorem.

Markov's work on the MLSq has been wrongly discussed. Neyman (1934) attributed to him Gauss' second justification of 1823 which even until our time (Dodge 2003, p. 161) is sometimes called after both Gauss and Markov. David & Neyman (1938) repeated the latter's mistake but the same year Neyman (1938b/1952, p. 228) corrected himself.

Then, Linnik et al (1951, p. 637) maintained that Markov had *in essence* introduced concepts *identical* to the modern notions of unbiased and effective statistics. Without explaining that latter notion I simply note that these authors should have replaced Markov by Gauss.

Markov (1899) upheld the second justification of the MLSq perhaps much more resolutely than his predecessors (the first such opinion appeared in 1825). However, he (1899/1951, p. 246) depreciated himself:

I consider [that justification] *rational since it does not obscure the conjectural essence of the method.* […] *We do not ascribe the ability* *of providing the most probable or the most plausible results to the method …*

Such a method does not need any justification. Furthermore, the MLSq does have optimal properties (Petrov 1954, cited above as well). Also see Sheynin (2006).

6.4. The Minimax Method

There also exist other methods of solving systems (6.1). They do not lead to the useful properties of the MLSq estimators but are expedient in some cases. And there also exists a special method leading to the least absolute value of the maximal residual

$$|v_{\max}| = \min. \qquad (6.4)$$

Least means least among all possible solutions (and therefore sets of v_i's); in the simplest case, among several reasonable solutions. The minimax method does not belong to probability theory, does not lead to any *best* results, but it allows to make definite conclusions. Recall that the coefficients a_i, b_i, … in system (6.1) are given by the appropriate theory and ask yourselves: do the observations w_i confirm it? After determining $\hat{x}, \hat{y},...$ (this notation does not infer the MLSq anymore) we may calculate the residual free terms v_i and determine whether they are not too large as compared with the known order of errors. In such cases we ought to decide whether the theory was wrong or that the observations were substandard. And here the minimax method is important: if even condition (6.4) leads to inadmissible $|v_i|$, our doubts are certainly justified.

Both Euler and Laplace had applied the minimax method (the latter had devised an appropriate algorithm) for establishing whether the accomplished meridian arc measurements denied the ellipticity of the figure of the Earth. Kepler (1609/1992, p. 334/143) could have well applied elements of that method for justifying his rejection of the Ptolemaic system of the world: the Tychonian observations were sufficiently precise but did not agree with it. In astronomy, equations are neither linear, nor even algebraic, and Kepler had to surmount additional difficulties (irrespective of the method of their solution).

Condition (6.4) is identical to having

$$\lim(v_1^{2k} + v_2^{2k} + ... + v_n^{2k}) = \min,\ k \to \infty,$$

which is almost obvious. Indeed, suppose that $|v_1|$ is maximal. Then, as $k \to \infty$, all the other terms of the sum can be neglected. For arriving at a minimal value of the sum, v_1^{2k}, and therefore $|v_1|$ also, should be as small as possible.

Without looking before he leapt, Stigler(1986, pp. 27, 28) confidently declared that Euler's work (see above) was *a statistical failure* since he

Distrusted the combination of equations, taking the mathematician's view that errors actually increase with aggregation rather than taking the statistician's view that random errors tend to cancel one another.

However, at the turn of the 18th century Laplace, Legendre and other scientists *refusa de compenser* les angles of a triangulation chain between two bases. Being afraid of corrupting the measured angles by adjusting them, they resolved to calculate each half of the chain from its own base (and somehow to adjust the common side of both parts of the triangulation), see Méchain & Delambre (1810, pp. 415 – 433). Later, in the Third Supplement to his *Théorie analytique*, Laplace (ca. 1819/1886, pp. 590 – 591) explained that decision by the lack of the *vraie théorie* which he (rather than Gauss whom he had not mentioned) had since created. See also Sheynin (1993a, p. 50).

In the Soviet Union, separate triangulation chains were included in the general adjustment of the entire network only after preliminarily adjustment (§ 1.1.4). This pattern was necessary since otherwise the work would have been impossible. In addition, the influence of the systematic errors should have been restricted to separate chains (as stated in a lecture of ca. 1950 of an eminent Soviet geodesist, A. A. Isotov), and this consideration was akin to the decision of the French scientists described above.

Chapter 7. Theory of Probability, Statistics, Theory of Errors

7.1. Axiomatization of the Theory of Probability

Following many previous author, I noted (§ 1.1.1) that the classical definition of probability is unsatisfactory and that Hilbert (1901/1970, p. 306) recommended to axiomatize the theory of probability:

Durch die Untersuchungen über die Grundlagen der Geometrie wird uns die Aufgabe nahe gelegt, nach diesem Vorbilds diejenigen physikalischen Disziplinen axiomatisch zu behandeln, in denen schon heute die Mathematik eine hervorragende Rolle spielt: dies sind in erster Linie die Wahrscheinlichkeitsrechnung und die Mechanik.

The theory of probability had then still been an applied mathematical (but not physical) discipline. In the next lines of his report Hilbert mentioned the method of mean values. That method or theory had been an intermediate entity divided between statistics and the theory of errors, and Hilbert was one of the last scholars (the last one?) to mention it, see Sheynin (2007, pp. 44 – 46).

Boole (1854/1952, p. 288) indirectly forestalled Hilbert:

The claim to rank among the pure sciences must rest upon the degree in which it [the theory of probability] *satisfies the following conditions: 1° That the principles upon which its methods are founded should be of an axiomatic nature.*

Boole formulated two more conditions, both of a general scientific nature. Attempts to axiomatize the theory began almost at once after Hilbert's report. However, as generally recognized, only Kolmogorov attained quite satisfactory results. Without discussing the essence of his work (see for example Gnedenko 1954, § 8 in chapter 1), I quote his general statements (1933, pp. III and 1):

Der leitende Gedanke des Verfassers war dabei, die Grundbegriffe der Wahrscheinlichkeitsrechnung, welche noch unlängst für ganz eigenartig galten, natürlicherweise in die Reihe der allgemeinen Begriffsbildungen der modernen Mathematik einzuordnen.

Die Wahrscheinlichkeitstheorie als mathematische Disziplin soll und kann genau in demselben Sinne axiomatisiert werden wie die Geometrie oder die Algebra. Das bedeutet, daß, nachdem die Namen der zu untersuchenden Gegenstände und ihrer Grundbeziehungen sowie die Axiome, denen diese Grundbeziehungen zu gehorchen haben, angegeben sind, die ganze weitere Darstellung sich ausschließlich auf diese Axiome gründen soll und keine Rücksicht auf die jeweilige konkrete Bedeutung dieser Gegenstände und Beziehungen nehmen darf.

For a long time these ideas had not been generally recognized (Doob 1989; 1994, p. 593):

To most mathematicians mathematical probability was to mathematics as black marketing to marketing; […] *The confusion between probability and the phenomena to which it is applied* […] *still plagues the subject;* [the significance of the Kolmogorov monograph] *was not appreciated for years, and some mathematicians sneered that* […] *perhaps probability needed rigor, but surely not rigor mortis;* […] *The role of measure theory in probability* […] *still embarrasses some who like to think that mathematical probability is not a part of analysis.*

It was some time before Kolmogorov's basis was accepted by probabilists. The idea that a (mathematical) random variable is simply a function, with no romantic connotation, seemed rather humiliating to some probabilists …

For a long time Hausdorff's merits had remained barely known. His treatise on the set theory (1914, pp. 416 – 417) included references to probability, but much more was contained in his manuscripts, see Girlich (1996) and Hausdorff (2006). I also mention Markov (1924). On p. 10 he stated a curious axiom and on p. 24 referred to it (without really thinking how the readers will manage to find it):

Axiom. [Not separated from general text!] *If* […] *events p, q, r, …, u, v are equally possible and divided with respect to event A into favourable and unfavourable, then*, [if] *A has occurred,* [those] *which are unfavourable to event A fall through, whereas the others remain equally possible …*

The addition and multiplication theorems along with the axiom mentioned above serve as an unshakeable base for the calculus of probability as a chapter of pure mathematics.

His axiom and statement have been happily forgotten.

Shafer & Vovk (2001) offered their own axiomatization, possibly very interesting but demanding some financial knowledge. They (2003, p. 27) had explained their aim:

[In our book] *we show how the classical core of probability theory can be based directly on game-theoretic martingales, with no appeal to measure theory. Probability again becomes* [a] *secondary concept but is now defined in terms of martingales …*

Barone & Novikoff (1978) and Hochkirchen (1999) described the history of the axiomatization of probability. The latter highly estimated an unpublished lecture of Hausdorff read in 1923.

7.2. Definitions and Explanations

As understood nowadays, statistics originated in political arithmetic (Petty, Graunt, mid-17th century). It quantitatively (rather than qualitatively) studied population, economics and trade, discussed the appropriate causes and connections and applied simplest stochastic considerations. Here is a confirmation (Kendall 1960):

Statistics, as we now understand the term, did not commence until the 17th century, and then not in the field of 'statistics' [Staatswissenschaft]. *The true ancestor of modern statistics is* […] *Political Arithmetic*.

Statistics had gradually, and especially since the second half of the 19th century, begun to penetrate various branches of natural sciences. This led to the appearance of the term *statistical method* although we prefer to isolate three stages of its development.

At first, conclusions were being based on (statistically) noticed qualitative regularities, a practice which conformed to the qualitative essence of ancient science. See the statements of Hippocrates (§ 3.1) and Celsus (§ 1.1.3).

The second stage (Tycho in astronomy, Graunt in demography and medical statistics) was distinguished by the availability of statistical data. Scientists had then been arriving at important conclusions either by means of simple stochastic ideas and methods or even directly, as before. A remarkable example is the finding of an English physician Snow (1855/1965, pp. 58 – 59) who compared mortality from cholera for two groups of the London population, of those whose drinking water was (somehow) purified or not. Purification decreased mortality by 8 times and he thus discovered the way in which cholera epidemics had been spreading.

During the present stage, which dates back to the end of the 19th century, inferences are being checked by quantitative stochastic rules.

The questions listed by Moses (Numbers 13:17 – 20) can also be attributed to that first stage (and to political arithmetic): he sent scouts to *spy out* the land of Canaan, to find out

whether the people who dwell in it are strong or weak, whether they are few or many, […] *whether the land is rich or poor …*

In statistics itself, exploratory data analysis was isolated. Already Quetelet discussed its elements (1846); actually, however, the introduction of isolines was a most interesting example of such analysis. Humboldt (1817, p. 466) invented isotherms and (much later) mentioned Halley who, in 1791, had shown isolines of magnetic declination over North Atlantic.

That analysis belongs to the scientific method at large rather than mathematics and is not therefore recognized by mathematical statistics. It only belongs to theoretical statistics which in my opinion should mostly explain the difference between the two statistical sisters. Some authors only recognize either one or another of them. In 1953 Kolmogorov (Anonymous 1954, p. 47), for example, declared that

We have for a long time cultivated a wrong belief in the existence, in addition to mathematical statistics and statistics as a socio-economic science, of something like yet another non-mathematical, although universal general theory of statistics which essentially comes to *mathematical statistics and some technical methods of collecting and treating statistical data. Accordingly, mathematical statistics was declared a part of this* <u>*general theory of statistics*</u>. *Such views* […] *are wrong.* […]

All that which is common in the statistical methodology of the natural and social sciences, all that which is here indifferent to the specific character of natural or social phenomena, belongs to […] *mathematical statistics*.

These *technical methods* indeed constitute exploratory data analysis.

Kolmogorov & Prokhorov (1974/1977, p. 721) defined mathematical statistics as

the branch of mathematics devoted to the mathematical methods for the systematization, analysis and use of statistical data for the drawing of scientific and practical inferences.

Recall (§ 2.7) that they also defined the notion of statistical data and note that a similar definition of the theory of statistics had appeared in the beginning of the 19th century (Butte 1808, p. XI): it is

Die Wissenschaft der Kunst statistische Data zu erkennen und zu würdigen, solche zu sammeln und zu ordnen.

The term *mathematical statistics* appeared in the mid-19th century (Knies 1850, p. 163; Vernadsky 1852, p. 237), and even before Butte Schlözer (1804) mentioned the theory of statistics in the title of his book. He (p. 86) also illustrated the term *statistics*: *Geschichte ist eine fortlaufende Statistik, und Statistik stillstehende Geschichte*. Obodovsky (1839, p. 48) offered a similar statement: *history is related to statistics as poetry to painting*.

Unlike Shlözer, many statisticians understood his pithy saying as a *definition* of statistics; as well we may say today: a car is a landed plane, and a plane, a car taken wing.

For us, the theory of statistics essentially originated with Fisher. A queer episode is connected here with Chuprov's book (1909/1959). Its title is *Essays on the Theory of Statistics*, but on p. 20 he stated that *A clear and strict justification of the statistical science is still needed*!

The determinate theory of errors (§ 6.1) has much in common with both the exploratory data analysis and Fisher's creation, the experimental design (a rational organization of measurements corrupted by random errors). However, the entire theory of errors seems to be the application of the statistical method to the process of measurement and observation in experimental science rather than a chapter of mathematical statistics, as it is usually maintained. Indeed, stellar statistics is the application of the statistical method to astronomy, and medical statistics is etc. Furthermore, unlike mathematical statistics the theory of errors cannot *at all* give up the notion of true value of a measured constant (§ 6.2).

7.3. Penetration of the Theory of Probability into Statistics

Hardly anyone will deny nowadays that statistics is based on the theory of probability, but the situation had not always been the same. Already Jakob Bernoulli (§ 4.1.1) firmly justified the possibility of applying the latter but statisticians had not at all been quick to avail themselves of the new opportunity. In those times, this might have been partially due to the unreliability of data; the main problem was their general treatment. Then, statistical investigations are not reduced to mathematical calculations; circumstances accompanying the studied phenomena are also important, cf. Leibniz' opinion in § 4.1.1. Finally, their education did not prepare statisticians for grasping mathematical ideas and perhaps up to the 1870s they stubbornly held to *equally possible cases*, that is, to the theoretical probability.

Lack of such cases meant denial of the possibility to apply probability theory. But forget the 1870s! In 1916 A. A. Kaufman (Ploshko & Eliseeva 1990, p. 133) declared that the theory of probability is only applicable to independent trials with constant probability of *success* and certainly only when those equally possible cases existed.

Now, Quetelet. He had introduced mean inclinations to crime and marriage (although not for separate groups of population), but somehow statisticians did not for a long time understand that mean values ought not to be applied to individuals. As a consequence, by the end of his life and after his death (1874), mathematically ignorant statisticians went up in arms against those inclinations and probability in general (Rümelin 1867/1875, p. 25):

Wenn mir die Statistik sagt, daß ich im Laufe des nächsten Jahres mit einer Wahrscheinlichkeit von 1 zu 49 sterben, mit einer noch größeren Wahrscheinlichkeit schmerzliche Lücken in dem Kreis mir theurer Personen zu beklagen haben werde, so muß ich mich unter den Ernst dieser Wahrheit in Demuth beugen; wenn sie aber, auf ähnliche Durchschnittszahlen gestützt, mir sagen wollte, daß mit einer Wahrscheinlichkeit von 1 zu so und so viel [I shall commit a crime] *so dürfte ich ihr unbedenklich antworten: ne sutor ultra crepidam*! [Cobbler! Stick to your last!].

A healthy man could have just as well rejected the conclusions drawn from a life table (Chuprov 1909/1959, pp. 211– 212).

Lexis infused a fresh spirit into (population) statistics. His followers, Bortkiewicz, Chuprov, Bohlmann, Markov, founded the so-called *continental direction of statistics*. In England, Galton, and Pearson somewhat later created the Biometric school which had been statistically studying Darwinism. The editors of its journal, *Biometrika*, a *Journal for the Statistical Study of Biological Problems*, were Weldon (a biologist who died in 1906), Pearson and Davenport (an author of a book on biometry and several articles) *in consultation* with Galton. Here is its Editorial published in 1902, in the first issue of that journal:

The problem of evolution is a problem in statistics. […] *We must turn to the mathematics of large numbers, to the theory of mass phenomena, to interpret safely our observations*. […] *May we not ask how it came about that the founder of our modern theory of descent made so little appeal to statistics*? […] *The characteristic bent of Darwin's mind led him to establish the theory of descent without mathematical conceptions; even so Faraday's mind worked in the case of electro-magnetism. But as every idea of Faraday allows of mathematical definition, and demands mathematical analysis*, […] *so every idea of Darwin – variation, natural selection* […] *– seems at once to fit itself to mathematical definition and to demand statistical* *analysis.* […] T*he biologist, the mathematician and the statistician have hitherto had widely differentiated field of work.* […] *The day will come* […] *when we shall find mathematicians who are competent biologists, and biologists who are competent mathematicians …*

During many years the Biometric school had been keeping to empiricism (Chuprov 1918 – 1919, t. 2, pp. 132 – 133) and he and Fisher (1922, pp. 311 and 329n) both indicated that Pearson confused theoretical and empirical indicators. And here is Kolmogorov (1947, p. 63; 1948/2002, p. 68):

The modern period in the development of mathematical statistics began with the fundamental works of […] *Pearson, Student, Fisher* […]. *Only in the contributions of the English school did the application of probability theory to statistics cease to be a collection of separate isolated problems and become a general theory of statistical testing of stochastic hypotheses* (*of hypotheses about laws of distribution*) *and of statistical estimation of parameters of these laws.*

The main weakness of the [Biometric] *school* [in 1912] *were: 1. Rigorous results on the proximity of empirical sample characteristics to the theoretical ones existed only for independent trials. 2. Notions of the logical structure of the theory of probability, which underlies all the methods of mathematical statistics, remained at the level of eighteenth century results. 3. In spite of the immense work of compiling statistical tables* […], *in the intermediate cases between 'small' and 'large' samples their auxiliary techniques proved highly imperfect.*

I (2010) have collected many pronouncements about the Biometric school and Pearson; hardly known outside Russia was Bernstein's high opinion. I note that Kolmogorov passed over in silence the Continental direction of statistics. Chuprov had exerted serious efforts for bringing together that Continental direction and the Biometric school, but I am not sure that he had attained real success. And this I say in spite of the Resolution of condolence passed by the Royal Statistical Society after the death of its Honorary member (Sheynin 1990/2011, p. 156). It stated that Chuprov's contributions (not special efforts!) *did much to harmonize the methods of statistical research developed by continental and British workers*. Even much later Bauer (1955, p. 26) reported that he had investigated how both schools had been applying analysis of variance and concluded (p. 40) that their work was going on side by side but did not tend to unification.

Bibliography

Abbreviation

AHES	= *Archive for History of Exact Sciences*
OC	= *Oeuvr. Compl.*

S, G, No. … = the source in question is translated on my website (either sheynin.de or google, oscar sheynin, home), see Document No. …

Anonymous (1948), *Vtoroe Vsesoiuznoe Soveshchanie po Matemeticheskoi Statistike* (Second All-Union Conference on Math. Statistics). Tashkent.

Anonymous (1954, in Russian), Survey of the scientific conference on issues of statistics. *Vestnik Statistiki*, No. 5, pp. 39 – 95.

Arbuthnot J. (1712), An argument for Divine Providence taken from the constant regularity observed in the birth of both sexes. In Kendall & Plackett (1977, pp. 30 – 34).

Aristotle (1908 – 1930, 1954), *Works*, vols 1 – 12. London. I am referring to many treatises from that source. There is also a new edition of Aristotle (Princeton, 1984, in two volumes).

Arnauld A., Nicole P. (1662), *L'art de penser.* Paris, 1992.

Barone J., Novikoff A. (1978), A history of the axiomatic formulation of probability from Borel to Kolmogorov. AHES, vol. 18, pp. 123 – 190.

Baer K. (1873), *Zum Streit über den Darwinismus*. Dorpat [Tartu].

Bayes T. (1764 – 1765), An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, vol. 53 for 1763, pp. 360 – 418; vol. 54 for 1764, pp. 296 – 325. Reprint of pt. 1: *Biometrika*, vol. 45, 1958, pp. 293 – 315, also Pearson & Kendall (1970, pp. 131 – 153). German translation of both parts: Leipzig, 1908.

Bauer R. K. (1955), Die Lexische Dispersionstheorie in ihren Beziehungen zur modernen statistischen Methodenlehre etc. *Mitteilungsbl. f. math. Statistik u. ihre Anwendungsgebiete*, Bd. 7, pp. 25 – 45.

Bernoulli D. (1738; 1982, pp. 223 – 234, in Latin), Exposition of a new theory on the measurement of risk. *Econometrica*, vol. 22, 1954, pp. 23 – 36.

--- (1770; 1982, pp. 306 – 324), Disquisitiones analyticae de nouo problemata coniecturale.

--- (1778; 1982, pp. 361 – 375, in Latin), The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. *Biometrika*, vol. 48, 1961, pp. 3 – 13, with translation of Euler (1778). Reprint: E. S. Pearson & Kendall (1970, pp. 155 – 172).

--- (1780; 1982, pp. 376 – 390), Specimen philosophicum de compensationibus horologicis, et veriori mensura temporis.

--- (1982), *Werke*, Bd. 2. Basel.

Bernoulli Jakob (1713), *Ars conjectandi.* Reprint: Bernoulli J. (1975, pp. 107 – 259).

--- (1975), *Werke*, Bd. 3. Basel. Includes reprints of several memoirs of other authors and commentaries.

--- (2005), *On the Law of Large Numbers*. Berlin, this being a translation of pt. 4 of the *Ars Conjectandi.* S, G,

Bernoulli Nikolaus (1709), *De Usu Artis Conjectandi in Jure*. Reprint: Bernoulli J. (1975, pp. 289 – 326).

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

--- (1946), *Teoria Veroiatnostei* (Theory of Prob.). Moscow – Leningrad. 4-th edition.

Bertrand J. (1888), *Calcul des probabilités.* 2nd ed., 1907. Reprints: New York, 1970, 1972. Second edition practically coincides with the first one.

Bervi N. V. (1899, in Russian), Determining the most probable value of the measured object independently from the Gauss postulate. *Izvestia Imp. Mosk. Obshchestvo Liubitelei Estestvoznania, Antropologii i Etnografii*, Sect. phys. sci., vol. 10, No. 1, pp. 41 – 45.

Bessel F. W. (1816), Untersuchungen über die Bahn des Olbersschen Kometen. *Abh. Preuss. Akad. Berlin*, math. Kl. 1812 – 1813, pp. 119 – 160. Bessel (1876) only contains a passage from this contribution.

--- (1818), *Fundamenta Astronomiae*. Königsberg. Fragment in Schneider I., Editor (1988), *Entwicklung der Wahrscheinlichkeitstheorie* etc. Darmstadt, pp. 277 – 279.

--- (1838), Untersuchung über die Wahrscheinlichkeit der Beobachtungsfehler. In Bessel (1876, Bd. 2, pp. 372 – 391).

--- (1839), Einfluß der Schwere auf die Figur eines … auflegenden Stabes. *Abh.*, Bd. 3, pp. 275 – 282.

--- (1876), *Abhandlungen*, Bde 1 – 3. Leipzig.

Bienaymé I. J. (1853), Considérations à l'appui de la découverte de Laplace etc. *C. r. Acad. Sci. Paris*, t. 37, pp. 309 – 324. Reprint: *J. Math. Pures et Appl.*, sér. 2, t. 12, 1867, pp. 158 – 176.

Boltzmann L. (1868), Studien über das Gleichgewicht der lebenden Kraft. In Boltzmann (1909, Bd. 1, pp. 49 – 96).

--- (1909), *Wissenschaftliche Abhandlungen*, Bde 1 – 3. Leipzig.

Bomford G. (1971), *Geodesy.* Oxford. First two editions: 1952, 1962.

Boole G. (1851), On the theory of probabilities. In author's book (1952, pp. 247 – 259).
--- (1854), On the conditions by which the solution of questions in the theory of probabilities are limited. Ibidem, pp. 280 – 288.
--- (1952), *Studies in Logic and Probability*, vol. 1. London.
Bortkiewicz L. von (Bortkevich V. I.) (1898a), *Das Gesetz der kleinen Zahlen.* Leipzig.
Boscovich R. (1758). *Philosophiae Naturalis Theoria.* Latin – English edition: Chicago – London, 1922. English translation from the edition of 1763: *Theory of Natural Philosophy.* Cambridge, Mass., 1966.
Buffon G. L. L. (1777), Essai d'arithmétique morale. In Buffon (1954, pp. 456 – 488).
--- (1954), *Œuvres philosophiques.* Paris. Editors, J. Piveteau, M. Fréchet, C. Bruneau.
Butte W. (1808), *Die Statistik als Wissenschaft.* Landshut.
Celsus A. C. (1935), *De Medicina*, vol. 1. London. In English. Written in 1st century.
Chebyshev P. L. (1845, in Russian), An essay on an elementary analysis of the theory of probability. In Chebyshev (1944 – 1951, vol. 5, pp. 26 – 87).
--- (1867), Des valeurs moyennes. *J. Math. Pures et Appl.*, t. 12, pp. 177 – 184.
--- (Lectures 1879/1880), *Teoria Veroiatnostei* (Theory of Probability). Moscow – Leningrad, 1936. **S, G,** No. 3.
--- (1899 – 1907), *Oeuvres*, tt. 1 – 2. Pétersbourg. Reprint: New York, 1962.
--- (1944 – 1951), *Polnoe Sobranie Sochineniy* (Complete Works), vols 1 – 5. Moscow – Leningrad.
Chuprov A. A. (1909), *Ocherki po Teorii Statistiki* (Essays on the Theory of Statistics). Moscow, 1959. Second ed., 1910
--- (1918 – 1919), Zur Theorie der Stabilität statistischer Reihen. *Skand. Aktuarietidskr.*, t. 1 – 2, pp. 199 – 256, 80 – 133.
--- (1960), *Voprosy Statistiki* (Issues in Statistics). Reprints and/or translations of papers. Moscow, 1960.
Clausius R. (1858), Über die mittlere Länge der Wege […] bei der Molekularbewegung. *Abhandlungen über die mechanische Wärmetheorie*, Abt. 2. Braunschweig, 1867, pp. 260 – 276.
Condorcet M. J. A. N. (1784), Sur le calcul des probabilités. *Hist. Acad. Roy. Sci. Paris 1781 avec Mém. Math. et Phys. pour la même année*. Paris, 1784, pp. 707 – 728.
Cournot A. A. (1843). *Exposition de la théorie des chances et des probabilités.* Paris, 1984. Editor B. Bru. **S, G,** No. 54.
D'Alembert J. Le Rond (1767), Doutes et questions sur le calcul des probabilités. In author's book *Mélanges de litterature, d'histoire et de philosophie*, t. 5. Amsterdam, pp. 239 – 264.
Danilevsky N. Ya. (1885), *Darvinism* (Darwinism), vol. 1, pts 1 – 2. Petersburg.
Darwin C. (1859), *Origin of Species*. London – New York, 1958. [Manchester, 1995.]
David F. N. (1962), *Games, Gods and Gambling.* London.
David F. N., Neyman, J. (1938), Extension of the Markoff theorem on least squares. *Stat. Res. Memoirs*, vol. 2, pp. 105 – 117.
De Moivre A. (1711, in Latin). De mensura sortis or the measurement of chance. *Intern.Stat. Rev.*, vol. 52, 1984, pp. 236 – 262. Commentary (A. Hald): Ibidem, pp. 229 – 236.
--- (1718), *Doctrine of Chances.* Later editions: 1738, 1756. References in text to reprint of last edition: New York, 1967.
--- (1725), *Treatise on Annuities on lives*. London. Two later editions: 1743 and 1756, incorporated in the last edition of his *Doctrine*, pp. 261 – 328. On p. xi of the same source (*Doctrine* 1756), the anonymous Editor stated that the appended *Treatise* was its improved edition [of 1743].
--- (1730), *Miscellanea Analytica de Seriebus et Quadraturis.* London.
--- (1733, in Latin), Transl. by author: A method of approximating the sum of the terms of the binomial $(a + b)^n$ expanded into a series from whence are deduced some practical rules to estimate the degree of assent which is to be given to experiments.

Incorporated in subsequent editions of the *Doctrine* (in 1756, an extended version, pp. 243 – 254).

--- (1756), This being the last edition of the *Doctrine*.

Dodge Y., Editor (2003), *Oxford Dictionary of Statistical Terms*. Oxford.

Doob J. L. (1989), Commentary on probability. In *Centenary of Math. in America*, pt. 2. Providence, Rhode Island, 1989, pp. 353 – 354. Editors P. Duren et al.

--- (1994), The development of rigor in mathematical probability, 1900 – 1950. *Amer. Math. Monthly*, vol. 103, 1996, pp. 586 – 595.

Dutka J. (1988), On the St. Petersburg paradox. AHES, vol. 39, pp. 13 – 39.

Eddington A. S. (1933), Notes on the method of least squares. *Proc. Phys. Soc.*, vol. 45, pp. 271 – 287.

Ehrenfest P. & T. (1907), Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. In Ehrenfest, P. (1959), *Coll. Scient. Papers*. Amsterdam, pp. 146 – 149.

Eisenhart C. (1963), Realistic evaluation of the precision and accuracy of instrument calibration. In Ku (1969, pp. 21 – 47).

Fisher R. A. (1922), On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.*, vol. A222, pp. 309 – 368.

Fourier J. B. J., (1826), Sur les résultats moyens déduits d'un grand nombre d'observations. *Œuvres*, t. 2. Paris, 1890, pp. 525 – 545.

Freudenthal H. (1951), Das Peterburger Problem in Hinblick auf Grenzwertsätze der Wahrscheinlichkeitsrechnung. *Math. Nachr.*, Bd. 4, pp. 184 – 192.

Galen C. (1951), *Hygiene.* Springfield, Illinois. Written in 2nd century.

Galton F. (1877), Typical laws of heredity. *Nature*, vol. 15, pp. 492 – 495, 512 – 514, 532 – 533. Also *Roy. Institution of Gr. Britain*, 1879, vol. 8, pp. 282 – 301.

--- (1889), *Natural Inheritance*. New York, 1973.

Gauss C. F. (1809, in Latin), *Theorie der Bewegung*, Book 2, Section 3. German translation in Gauß (1887), pp. 92 – 117.

--- (1816), Bestimmung der Genauigkeit der Beobachtungen. Ibidem, pp. 129 – 138.

--- (1821 – 1823), Preliminary author's report about Gauss (1823, pt. 1 – 2). Ibidem, pp. 190 – 199.

--- (1823, in Latin), Theorie der den kleinsten Fehlern unterworfenen Combination der Beobachtungen, pts 1 – 2. Ibidem, pp. 1 – 53.

--- (1828), Supplement to Gauss (1823). Ibidem, pp. 54 – 91.

--- (1863 – 1930), *Werke*, Bde 1 – 12. Göttingen a.o. Reprint: Hildesheim, 1973 – 1981.

--- (1887), *Abhandlungen zur Methode der kleinsten Quadrate*. Hrsg, A. Börsch & P. Simon. Latest ed.: Vaduz, 1998.

--- (1995, in Latin and English), Theory of combinations of observations etc. Includes author's preliminary report about that contribution. Translated with Afterword by G. W. Stewart. Philadelphia.

Gini, C. (1946), Gedanken von Theorem von Bernoulli. *Z. für Volkswirtschaft u. Statistik*, 82. Jg., pp. 401 – 413.

Girlich H.-J. (1996), Hausdorffs Beiträge zur Wahrscheinlichkeitstheorie. In Brieskorn E., Editor *Felix Hausdorff zum Gedächtnis*, Bd. 1. Braunschweig, pp. 31 – 70.

Gnedenko B. V. (1954, in Russian). *Theory of probability.* Moscow, 1973. [Providence, RI, 2005.] First Russian edition, 1950.

Great Books (1952), *Great Books of the Western World*, vols 1 – 54. Chicago.

Hald A. (1990), *History of Probability and Statistics and Their Applications before 1750.* New York.

--- (1998), *History of Mathematical Statistics from 1750 to 1930.* New York.

Hausdorff F. (1914), *Grundlehren der Mengenlehre*. Leipzig.

--- (2006), *Ges. Werke*, Bd. 5. Berlin.

Helmert F. R. (1904), Zur Ableitung der Formel von Gauss für die mittleren Beobachtungsfehler und ihrer Genauigkeit. *Z. f. Vermessungswesen*, Bd. 33, pp. 577 – 587. Also *Sitz.-Ber. Kgl. Preuss. Akad. Wiss. Berlin*, 1904, Hlbbd 1, pp. 950 – 964.

Herschel W. (1817), Astronomical observations and experiments tending to investigate the local arrangement of celestial bodies in space. Ibidem, pp. 575 – 591.

--- (1912), *Scientific papers*, vols. 1 – 2. London. [London, 2003.]

Hilbert D. (1901), Mathematische Probleme. *Ges. Abh.*, Bd. 3. Berlin, 1970, pp. 290 – 329. Mathematical problems. *Bull. Amer. Math. Soc.*, vol. 8, No. 10, 1902, pp. 403 – 479.

Hippocrates (1952), Aphorisms. In *Great Books* (1952, vol. 10, pp. 131 – 144).

Hochkirchen C. (1999), *Die Axiomatisierung der Wahrscheinlichkeitsrechnung.* Göttingen.

Humboldt A. (1817), Des lignes isothermes. *Mém. Phys. Chim. Soc. d'Arcueil*, t. 3, pp. 462 – 602.

Huygens C. (1657), De calcul dans les jeux de hasard. In Huygens (1888 – 1950, t. 14, 1920, pp. 49 – 91).

--- (1669, in French), Correspondence. In Huygens (1888 – 1950, t. 6. 1895, 515 – 518, 524 – 532, 537 – 539.

--- (1888 – 1950), *Oeuvres complètes*, tt. 1 – 22. La Haye.

Kamke E. (1933), Über neuere Begründungen der Wahrscheinlichkeitsrechnung. *Jahresber. Deutschen Mathematiker-Vereinigung*, Bd. 42, pp. 14 – 27.

Kendall M. G. (Sir Maurice) (1960), Where shall the history of statistics begin? *Biometrika*, vol. 47, pp. 447 – 449. Reprint: Ibidem, pp. 45 – 46.

Kendall M. G., Plackett R. L., Editors (1977), *Studies in the History of Statistics and Probability*, vol. 2. London. Collected reprints.

Kepler J. (1596, in Latin). Mysterium cosmographicum. *Ges. Werke*, Bd. 8. Münich, 1963, pp. 7 – 128, this being a reprint of second edition (1621, with additions to many chapters). English translation: New York, 1981.

--- (1601, in Latin), On the most certain fundamentals of astrology. *Proc. Amer. Phil. Soc.*, vol. 123, 1979, pp. 85 – 116.

--- (1604, in German), Thorough description of an extraordinary new star. *Vistas in Astron.*, vol. 20, 1977, pp. 333 – 339.

--- (1606, in Latin), *Über den neuen Stern im Fuß des Schlangenträger*. Würzburg, 2006.

--- (1609, in Latin), *New Astronomy.* Cambridge, 1992. Double paging.

--- (1610), Tertius interveniens. Das ist Warnung an etliche Theologos, Medicos und Philosophos. *Ges. Werke*, Bd. 4. München, 1941, pp. 149 – 258.

--- (1619, in Latin), *Welt-Harmonik*. München – Berlin, 1939. English translation: *Harmony of the World.* Philadelphia, 1997.

--- (1620 – 1621, in Latin), *Epitome of Copernican Astronomy*, Books 4 and 5. *Great Books* (1952, vol. 16, pp. 845 – 1004).

Khinchin A. Ya. (1927), Über das Gesetz der großen Zahlen. *Math. Ann.*, Bd. 96, pp. 152 – 168.

--- (1961, in Russian), The Mises frequency theory and modern ideas of the theory of probability. *Science in Context*, vol. 17, 2004, pp. 391 – 422.

Knies C. G. A. (1850), *Die Statistik als selbstständige Wissenschaft.* Kassel.

Kohli K. (1975b), Aus dem Briefwechsel zwischen Leibniz und Jakob Bernoulli. In Bernoulli J. (1975, pp. 509 – 513).

Kolmogorov A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin. *Foundations of the Theory of Probability.* New York, 1956.

--- (1947, in Russian), The role of Russian science in the development of the theory of probability. *Uch. Zapiski Moskovsk. Gos. Univ* No. 91, pp. 53 – 64. **S, G,** No. 47.

--- (1948, in Russian), The main problems of theoretical statistics. Abstract. In Anonymous (1948, pp. 216 – 220). S, G, No. 6.

--- (1954, in Russian), Law of small numbers. *Bolshaia Sov. Enz.* (Great Sov. Enc.), 2nd ed., vol. 26, p. 169. Published anonymously. **S, G,** No. 5.

--- (1963), On tables of random numbers. *Sankhya, Indian J. Stat.*, vol. A25, pp. 369 – 376.

--- (1985 – 1986, in Russian), *Selected Works*, vols. 1 – 2. Dordrecht, 1991 – 1992.

Kolmogorov A. N., Petrov A. A., Smirnov Yu. M. (1947, in Russian), A formula of Gauss in the method of least squares. In Kolmogorov (1992, pp. 303 – 308).

Kolmogorov A. N., Prokhorov Yu. V. (1974, in Russian), Mathematical statistics. *Great Sov. Enc.*, 3rd edition, vol. 15, pp. 480 – 484. This edition of the G. S. E. was translated into English; vol. 15 appeared in 1977.

Ku H. H., Editor (1969), *Precision Measurement and Calibration.* Nat. Bureau Standards Sp. Publ. 300, vol. 1. Washington.

Lambert J. H. (1760), *Photometria.* Augsburg.

--- (1765), Anmerkungen und Zusätze zur practischen Geometrie. In author's book *Beyträge zum Gebrauche der Mathematik und deren Anwendung*, Tl. 1. Berlin, 1765, pp. 1 – 313.

--- (1771), *Anlage zur Architektonik*, Bd. 1. *Phil. Schriften*, Bd. 3. Hildesheim, 1965.

--- (1772 – 1775), Essai de la taxéometrie ou sur la mesure de l'ordre. Ibidem, Bd. 8/1. Hildesheim, 2007, pp. 423 – 460.

Langevin P. (1913), La physique du discontinu. In collected articles *Les progrès de la physique moléculaire*. Paris, 1914, pp. 1 – 46.

Laplace P. S. (1776), Recherches sur l'intégration des équations différentielles aux différences finies. OC, t. 8. Paris, 1891, pp. 69 – 197.

--- (1781), Sur les probabilités. OC, t. 9. Paris, 1893, pp. 383 – 485.

--- (1786), Suite du mémoire sur les approximations des formules qui sont fonctions de très grands nombres. OC, t. 10. Paris, 1894, pp. 295 – 338.

--- (1796), *Exposition du système du monde.* OC, t. 6. Paris, 1884 (the whole volume, this being a reprint of the edition of 1835).

--- (1812), *Théorie analytique des probabilités.* OC, t. 7, No. 1 – 2. Paris, 1886. Consists of two parts, an Introduction (1814) and supplements, see below. Theory of probability proper is treated in pt. 2.

--- (1814), *Essai philosophique sur les probabilités.* In Laplace (1812/1886, No. 1, separate paging). *Philosophical Essay on Probabilities.* New York, 1995. Translator and editor A.I. Dale.

--- (1816), *Théor. anal. prob., Supplément 1.* OC, t. 7, No. 2, pp. 497 – 530.

--- (1818), *Théor. anal. prob., Supplément 2.* Ibidem, pp. 531 – 580.

--- (ca. 1819), *Théor. anal. prob., Supplément 3.* Ibidem, pp. 581 – 616.

Laurent H. (1873), *Traité du calcul des probabilités.* Paris.

Legendre A. M. (1805), *Nouvelles méthodes pour la détermination des orbites des comètes.* Paris.

--- (1814), Méthode des moindres quarrés. *Mém. Acad. Sci. Paris*, Cl. sci. math. et phys., t. 11, pt. 2, année 1910, pp. 149 – 154.

Lévy, P. (1925), *Calcul des probabilités.* Paris.

Linnik, Yu. V. (1951), Commentary on Markov (1916). In Markov (1951, pp. 668 – 670).

Linnik Yu. V. et al (1951), Survey of Markov's work in the number theory and the theory of probability. In Markov (1951, pp. 614 – 640). **S, G,** No. 5.

Maimonides M. (1975), *Introduction to the Talmud*. New York.

--- (1977), *Letters*. New York.

Markov A. A. (1899), The law of large numbers and the method of least squares. In Markov (1951, pp. 231 – 251). **S, G,** No. 5.

--- (1900), [*Treatise*] *Ischislenie Veroiatnostei* (Calculus of Probabilities). Subsequent editions: 1908, 1913, and (posthumous) 1924. German translation: Leipzig – Berlin, 1912.

--- (1911, in Russian), On the basic principles of the calculus of probability and on the law of large numbers. In Ondar (1977/1981, pp. 149 – 153).

--- (1914), On the Problem of Jakob Bernoulli. In Markov (1951, pp. 511 – 521).

--- (1916), On the coefficient of dispersion. In Markov (1951, pp. 523 – 535). **S, G,** No. 5.

--- (1951), *Izbrannye Trudy* (Sel. Works). N. p.

Maupertuis P. L. M. (1756), Sur le divination. *Oeuvres*, t. 2. Lyon, 1756, pp. 298 – 306.

Maxwell J. C. (1860), Illustrations of the dynamical theory of gases. In author's *Scient.Papers*, vol. 1. Paris, 1927, pp. 377 – 410.

Méchain P. F. A., Delambre J. B. J. (1810), *Base du système métrique décimale*. Paris, t. 3.

Mendeleev D. I. (1877), The oil industry of Pennsylvania and in the Caucasus. *Sochinenia* (Works). Moscow – Leningrad, vol. 10, 1949, pp. 17 – 244.

Michell J. (1767), An inquiry into the probable parallax and magnitude of the fixed stars. *Phil. Trans. Roy. Soc. Abridged*, vol. 12, 1809, pp. 423 – 438.

Mill J. S. (1843), *System of Logic.* London, 1886. Many more editions, e. g. *Coll. Works*, vol. 8. Toronto, 1974

Montmort P. R. (1708), *Essay d'analyse sur les jeux de hazard.* Second ed., 1713. Published anonymously. References in text to reprint: New York, 1980. **S, G,** No. 58 (Introduction).
Neugebauer O. (1950), The alleged Babylonian discovery of the precession of the equinoxes. In Neugebauer (1983, pp. 247 – 254).
--- (1983), *Astronomy and History. Sel. Essays.* New York.
Newcomb S. (1860), [Discussion of the principles of probability theory], *Proc. Amer. Acad. Arts and Sciences*, vol. 4 for 1857 – 1860, pp. 433 – 440.
--- (1886), A generalized theory of the combination of observations. *Amer. J. Math*., vol. 8, pp. 343 – 366.
Newton I. (1967), *Mathematical Papers*, vol. 1. Cambridge.
Neyman J. (1934), On two different aspects of the representative method. *J. Roy. Stat. Soc.*, vol. 97, pp. 558 – 625. Reprinted in Neyman (1967, pp. 98 – 141).
--- (1938a), L'estimation statistique traitée comme un problème classique de probabilité. Reprinted Ibidem, pp. 332 – 353.
--- (1938b), *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, 1952.
--- (1967), *Selection of Early Statistical Papers.* Berkeley.
Obodovsky A. (1839), *Teoria Statistiki* (Theory of Statistics). Petersburg.
Ondar Kh. O., Editor (1977, in Russian), *Correspondence between Markov and Chuprov on the Theory of Probability and Mathematical Statistics.* New York, 1981.
Oresme N. (1966), *De Proportionibus Proportionum* and *Ad Pauca Respicientes.* Editor E. Grant. Madison. Latin – English edition. Written in the 14th c.
Pascal B. (1998 – 2000), *Oeuvres complètes*, tt. 1 – 2. Paris.
Pearson E. S., Kendall M. G., Editors (1970), *Studies in the History of Statistics and Probability* [vol. 1]. London. Collection of reprints.
Pearson K. (1893), Asymptotical frequency curves. *Nature*, vol. 15, pp. 615 – 616.
--- (1905), Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A rejoinder. *Biometrika*, vol. 4, pp. 169 – 212.
--- (1924), Historical note on the origin of the normal curve of errors. *Biometrika*, vol. 16, pp. 402 – 404.
--- (1925), James Bernoulli's theorem. *Biometrika*, vol. 17, pp. 201 – 210.
--- (1978), *History of Statistics in the 17th and 18th Centuries against the Changing Background of Intellectual, Scientific and Religious Thought.* Lectures 1921 – 1933. Editor E. S. Pearson. London.
Peirce B. (1873), On the errors of observations. Appendix 2 to *Report of the Superintendent of US Coast Survey*. Washington, pp. 200 – 224.
Petrov V. V. (1954, in Russian), Method of least squares and its extreme properties. *Uspekhi Matematich. Nauk*, vol. 9, pp. 41 – 62.
Petruszewycz M. (1983), Description statistique de textes littéraires russes par la méthodes de Markov. *Rev. Études Slaves*, t. 55, pp. 105 – 113.
Petty W. (1662), A Treatise of Taxes and Contributions. In Petty (1899, vol. 1, pp. 1 – 97).
--- (1899), *Economic Writings*, vols 1 – 2. Ed., C. H. Hull. Cambridge. Reprint: London, 1997.
Ploshko B. G., Eliseeva I. I. (1990), *Istoria Statistiki* (History of Statistics). Moscow.
Poincaré H. (1896), *Calcul des probabilités.* Paris, 1912, reprinted 1923 and 1987.
--- (1921), Résumé analytique [of own works]. In *Mathematical Heritage of H. Poincaré.* Providence, Rhode Island, 1983. Editor F. E. Browder, pp. 257 – 357.
Poisson S. D. (1824), Sur la probabilité des résultats moyens des observations, pt. 1. *Conn. des Temps* for 1827, pp. 273 – 302.
--- (1825 – 1826), Sur l'avantage du banquier au jeu de trente-et-quarante. *Annales Math. Pures et Appl.*, t. 16, pp. 173 – 208.
--- (1837), *Recherches sur la probabilité des jugements* etc. Paris, 2003.
Polya G. (1920), Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Math. Z.*, Bd. 8, pp. 171 – 181.
Prokhorov Yu. V. Editor (1999), *Veroiatnost i Matematicheskaia Statistika. Enziklopedia* (Probability and Math. Statistics. Enc.). Moscow.

Prokhorov Yu. V., Sevastianov B. A. (1999), Probability theory. In Prokhorov (1999, pp. 77 – 81).
Quetelet A. (1846), *Lettres ... sur la théorie des probabilités.* Bruxelles.
--- (1853), *Théorie des probabilités.* Bruxelles.
Rabinovitch N. L. (1973), *Probability and Statistical Inference in Ancient and Medieval Jewish Literature.* Toronto.
Rümelin G. (1867), Über den Begriff eines socialen Gesetzes. In author's *Reden und Aufsätze*. Freiburg i/B – Tübingen, 1875, pp. 1 – 31.
Rumshitsky L. Z. (1966), *Elementy Teorii Veroiatnostei* (Elements of the Theory of Probability). Moscow.
Schlözer A. L. (1804), *Theorie der Statistik.* Göttingen.
Schmeidler F. (1984), *Leben und Werke des … Astronomen Bessel.* Kalkheim/T.
Seidel L. (1865), Über den … Zusammenhang … zwischen der Häufigkeit der Typhus-Erkrankungen und dem Stande des Grundwassers. *Z. Biol.*, Bd. 1, pp. 221 – 236.
--- (1866), Vergleichung der Schwankungen der Regenmengen mit der Schwankungen in der Häufigkeit des Typhus. Ibidem, Bd. 2, pp. 145 – 177.
Shafer G., Vovk V. (2001), *Probability and Finance.* New York.
Sheynin O. (1968), On the early history of the law of large numbers. *Biometrika*, vol. 55, pp. 459 – 467. Reprint: Pearson & Kendall (1970, pp. 231 – 239).
--- (1971a), Newton and the theory of probability. AHES, vol. 7, pp. 217 – 243.
--- (1971b), Lambert's work in probability. AHES, vol. 7, pp. 244 – 256.
--- (1977), Early history of the theory of probability. AHES, vol. 17, pp. 201 – 259.
--- (1990, in Russian), *Alexandr A. Chuprov: Life, Work, Correspondence*. V& R Unipress, 2011.
--- (1993a), On the history of the principle of least squares. AHES, vol. 46, pp. 39 – 54.
--- (1993b), Treatment of observations in early astronomy. AHES, vol. 46, pp. 153 – 192.
--- (1996), *History of the Theory of Errors*. Egelsbach.
--- (1999a), Gauss and the method of least squares. *Jahrbücher f. Nationalökonomie u. Statistik*, Bd. 219, pp. 458 – 467.
--- (1999b, in Russian), Slutsky: 50 years after his death. *Istoriko-Matematich. Issledovania*, No. 3 (38), pp. 128 – 137. **S, G,** No. 1.
--- (2000), Bessel: some remarks on his work. *Hist. Scientiarum*, vol. 10, pp. 77 – 83.
--- (2003), Geometric probability and the Bertrand paradox. Ibidem, vol. 13, pp. 42 – 53.
--- (2006), Markov's work on the treatment of observations. *Hist. Scientiarum*, vol. 16, pp. 80 – 95.
--- (2007), The true value of a measured constant and the theory of errors. *Hist. Scientiarum*, vol. 17, pp. 38 – 48.
--- (2008), Bortkiewicz' alleged discovery: the law of small numbers. *Hist. Scientiarum*, vol. 18, pp. 36 – 48.
--- (2010a), Karl Pearson: a century and a half after his birth. *Math. Scientist*, vol. 35, pp. 1 – 9.
--- (2010b), The inverse law of large numbers. Ibidem, pp. 132 – 133.
--- (2012), New exposition of Gauss' final justification of least squares. *Math. Scientist*, vol. 37, pp. 147 – 148.
Simpson J. Y. (1847 – 1848), Anaesthesia. *Works*, vol. 2. Edinburgh, 1871, pp. 1 – 288.
Simpson T. (1756), On the advantage of taking the mean of a number of observations. *Phil. Trans. Roy. Soc.*, vol. 49, pp. 82 – 93.
--- (1757), Extended version of same: in author's book *Miscellaneous Tracts on Some Curious* […] *Subjects* […]. London, pp. 64 – 75.
Smirnov N. W., Dunin-Barkovsky I. W. (1959, in Russian), *Mathematische Statistik in der Technik*. Berlin, 1969.
Snow J. (1855), On the mode of communication of cholera. In *Snow on Cholera.* New York, 1965, pp. 1 – 139.
Stigler S. M. (1983), Who discovered Bayes's theorem? In author's book *Statistics on the Table*. Cambridge, Mass. – London, 1999, pp. 291 – 301.
--- (1986), *History of Statistics.* Cambridge (Mass.).

--- (1999), *Statistics on the Table.* Cambridge (Mass.). Collected revised articles.
Todhunter I. (1865), *History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace.* New York, 1949, 1965.
Tutubalin V. N. (1972), *Teoria Veroiatnostei v Estestvoznanii* (Theory of probability in Natural Sciences.). Moscow. **S, G,** No. 45.
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 as *Russ. Math. Surveys*.
Vernadsky V. I. (1852), Aims of statistics. In Druzinin N. K. (1963), *Chrestomatia po Istorii Russkoi Statistiki* (Reader in the History of Russian Statistics). Moscow, pp. 221 – 238.
Ventzel Elena S. (1969), *Teoria veroiatnostei* (Theory of Probability). 4th edition. Moscow.
Vovk V. G., Shafer G. R. (2003), Kolmogorov's contributions to the foundations of probability. *Problems of Information Transmission*, vol. 39, pp. 21 – 31.
Youshkevitch A. P. (1986, in Russian), N. Bernoulli and the publication of J. Bernoulli's Ars Conjectandi. *Theory of Prob. and Its Applications*, vol. 31, 1987, pp. 286 – 303.

Index of Names

The numbers refer to subsections rather than pages.

L. Z. Rumshisky

Elementary Theory of Probability, third edition

Л. З. Румшиский, *Элементы теории вероятностей*. Москва, 1966

Annotation

This book is an educational aid for courses on probability theory read in many technical academic institutions. Written in compliance with the approved curriculum, it fills the existing gap between university courses too difficult for students of academic institutions and popular literature. For understanding this aid it suffices to master mathematical analysis as taught in those institutions. Apart from students, it can be useful for engineers, especially of machine-building specialities, radio engineers and economists.

Contents

Preface

This educational aid is intended for [students of] technical academic institutions whose mathematical curriculum includes elements of the theory of probability and mathematical statistics taught for not more than 30 hours. The reader is supposed to master mathematical analysis as usually demanded by those institutions. This aid describes the main notions and some methods of the theory of probability which are nowadays necessary in many branches of technology. Not discussed is the theory of random processes or some special issues requiring more serious mathematical knowledge. Some of the more difficult parts of this aid are provided in small print and can at first be omitted. The examples are essential for explaining the main notions and I recommend the readers to study them attentively.

The aid is based on my lectures reads for ten years at the Moscow power engineering institute. Acknowledgements are due to A. M. Yaglom for his inestimable advice and comments as well as to R. Ya. Berry, I. A. Brin, M. I. Vishik, S. A. Stebakov and R. S. Khasminsky for useful indications.

Introduction

In various branches of technology and manufacturing it is ever more needful to deal with mass phenomena having special inherent peculiarities. Thus, when machine parts are processed by an automatic lathe, their sizes fluctuate around some standard value. These fluctuations are random: the knowledge of the sizes of the finished parts will not enable us to predict precisely the size of the next part. However, the distribution of the sizes in a large batch reveals a rather precise regularity. Their arithmetical means in different batches are approximately identical, and deviations of a given magnitude of one or another size from their mean in different batches occur almost equally often. A similar regularity is observed when repeatedly weighing the same object on a precise balance. Here also separate results differ but the mean of many of them will remain practically invariable. The frequency of some deviation from that mean can be precisely calculated. Such regularities are certainly unable to predict a separate result but allow us to treat the outcome of mass measurements.

It is a special mathematical science, the theory of probability, that studies the regularities inherent in various mean characteristics, in the repetition of random deviations of a given magnitude [from the appropriate mean] etc.

Such regularities had been first revealed when solving problems in games of chance, especially in dice games in the 17th century [the author neglects card games!]: when repeatedly casting a die, each of the six outcomes occurs almost equally often, with relative frequency of ca. 1/6. When rolling two dice, the sum of the occurring points takes its unequally possible values from 2 to 12. However, in a large number of casts their relative frequencies will be close to certain numbers which can be calculated beforehand by simple rules (see § 2.2.1).

The establishment of such rules and the solution of somewhat more complicated problems connected with dice games had been important during the initial period of the development of probability theory. Even now some main notions of that theory (random event and its probability, random variable etc) are expediently illustrated by examples about dice games. The six outcomes of a cast of a die clearly elucidates the notion of a *trial* with some equally possible outcomes and the numbers of the occurring points provide a simple example of a random variable, of a magnitude whose values depend on chance.

However, the simplest patterns worked out for solving problems connected with those games are very restrictedly applicable. Various other problems, formulated in the 19th and 20th centuries by the developing technology and natural sciences [and even earlier by population statistics and insurance], required a study of random variables of an essentially more complicated nature and especially of continuous random variables (the size of a manufactured machine part and the result of weighing (see Preface).

This aid pays main attention to such random variables and their characteristics; accordingly, the initial notions of random event and probability are only briefly discussed here, in Chapter 1. The second Chapter deals with (chiefly one-dimensional) random variables of the two separately discussed most important types, discrete and continuous, and their functions.

The third Chapter treats the main numerical characteristics of random variables, their simplest properties are proved. The most important properties of mean values connected with the so-called law of large numbers are discussed in Chapter 4. The fifth is devoted to a central issue of probability theory, to the limiting theorems ascertaining the role of the so-called normal law of distribution (in particular, for estimating the mean values). Chapter 6 is concerned with some applications of the theory to the mathematical treatment of measurements (to the theory of errors), and, finally, the seventh Chapter considers the problem of linear correlation between random variables, an issue important for practical applications.

Chapter 1. Random Events and Probabilities

1.1. Random events. Frequency and probability

Random events are such that can occur or not after a certain set of conditions had been realized. These conditions are certainly supposed to be essentially connected with the possibility of the occurrence of these events and to be unboundedly reproducible. Each such reproduction is called a *trial*.

Example: a cast of a die. The random event is here the outcome of a six, of an even number of points etc. *Another example*: weighing an object. The random variable is the restriction of the error of that weighing: it does not exceed a number established beforehand [a very clumsy explanation]. The error is understood as the difference between the result obtained and the true value[1] of the object's weight.

Relative frequency of a random event is the ratio of the number m of its occurrences to the total number n of trials. Experience shows that *after numerous trials the frequency m/n of the random event possesses certain stability*. If, for example, after a large number of trials the frequency became $m/n = 0.2$, then, in any other set of a sufficiently large number n_1 of trials the frequency m_1/n_1 will be close to 0.2. Therefore, the frequencies of a certain random variable in such series of trials as though group themselves around some number. If, for example, a precise and homogeneous cube (a *regular* die) is cast, the frequencies of each possible outcome fluctuate around one and the same number, 1/6.

The stability of relative frequencies can only be explained as a manifestation of some objective property of random events. The fact just described is a corollary of the regularity of the die and therefore of the equally possible outcomes 1, 2, …, 6. *The degree of the objective possibility of the occurrence of a random event can be measured by a number*, the probability of the random event. *The relative frequencies of the occurrence of that random event are grouped around that very number*.

Both that frequency and the probability of a random event should be dimensionless magnitudes situated between 0 and 1. However, given the set of conditions, the former also depends on the executed trials whereas the probability of a random event is only connected with the conditions. Probability *is an initial main notion; in general, it cannot be defined by simpler notions*. It can only be directly calculated in some simplest patterns (§ 1.2). An analysis of such patterns allows us to establish the main properties of probability, necessary for continuing the description of our subject.

1.2. Classical definition of probability

First, some terminology. *Random events are called incompatible (exclusive) if they cannot occur at the same time. If one and only one of them ought to occur in every trial, they make up a complete group of pairwise incompatible events*[2]. In this section, we restrict our attention to *trials with equally possible outcomes* but will not explain the notion of *equal possibility* by simpler notions. It is usually justified by some consideration of symmetry (see the example with the casting of a die). In practice, it is connected with the equality of relative frequencies of all the outcomes in a large number of trials. In this section, the number of cases is always supposed finite.

More precisely, we suppose that *the outcomes of the trials can be represented as a complete group of pairwise incompatible and equally possible random events* or cases. If the complete group consists of N cases, each of them is assigned probability $1/N$. This assumption agrees with the fact that in a large number of trials equally possible cases occur almost equally often. In other words, those cases have relative frequencies close to $1/N$. In a cast of a regular die all the possible cases constitute a complete group and each has probability 1/6.

Consider now a compound event A, the occurrence of any of M fixed cases out of N possible ones. *By definition*, the probability of A, $P(A)$, is M/N. For example, the probability of the occurrence of an even number of points of a die is 3/6 since it only occurs in 3 cases out of 6. The formula

$$P(A) = M/N \qquad (1.1)$$

expresses the so-called *classical definition of probability*: *if the results of a trial can be represented as a complete group of N equally possible and pairwise incompatible cases and a random event A only occurs in M cases, its probability is M/N, the ratio of the number of cases favouring A to the total number of all cases*.

Example. Toss two coins and heads can appear twice, once, or not at all. Required are the probabilities of each of these three random events. For each coin the occurrence and non-occurrence of heads is supposed equally possible. The listed events constitute a complete group of obviously incompatible but not equally possible events. For applying formula (1.1) we ought to represent all the possible outcomes of the trial as a complete group of equally possible events. So, the cases are:

heads, heads; heads, tails; tails, heads; tails, tails

It is natural to consider that these *four* outcomes are equally possible, and once again they form a complete group of pairwise incompatible events. And now we may indeed apply the classical definition of probability. The appearance of 2, 1 and 0 heads have probabilities 1/4, 2/4 = 1/2 and 1/4.

I emphasize once more: the definition (1.1) is essentially based on the assumption of equal possibility of all the outcomes. All the problems to which this definition is applicable belong to the following simple pattern of *random sampling*: we randomly choose one element out of a set of *N* elements so that all of them have the same possibility of being selected. The event *A* is the choice of one of the *M* elements possessing a definite indication.

This pattern is easiest to imagine by introducing an urn. And so, an urn contains *N* balls identical to the touch, and only *M* of them are white. The trial consists in blindly extracting a ball, and the random event is the occurrence of a white ball. Its probability is *M/N*.

1.3. Main properties of probabilities. Addition of probabilities

An analysis of definition (1.1) allows us to reveal the following main properties of probabilities.

1) The probability of a random event is a non-negative number

$$P(A) \geq 0. \qquad (1.2)$$

2) A certain event, such that under a given set of conditions it ought to occur certainly, has probability

$$P(\text{certain event}) = 1. \qquad (1.3)$$

3) The probabilities of random events obey the addition rule. If event *C* consists in the occurrence of any one of two incompatible events *A* and *B*, its probability is the sum of their probabilities:

$$P(A + B) = P(A) + P(B). \qquad (1.4)$$

This equality also represents *the property of additivity of probabilities*.

The first two properties, (1.2) and (1.3), directly follow from (1.1) in which $M \geq 0$ and $N > 0$ and for a certain event $M = N$. The third property (1.4) for the pattern of random sampling is proved thus. Suppose that an urn contains N balls, K of them red, L, blue, and the rest are white. A ball is extracted and A and B are the occurrences of a red and a blue ball. Then the event $(A + B)$ is the appearance of a coloured ball. A direct calculation of probabilities by formula (1.1) provides

$P(A) = K/N$, $P(B) = L/N$, $P(A + B) = (K + L)/N$, QED.

It is extremely important that the properties of probabilities as described above hold not only for the pattern of random sampling but for any system of random events. Indeed, recall that we have established the general notion of probability by issuing from the stability of the relative frequencies of random events. It is therefore natural to suppose that the main properties of probabilities of random events coincide with those of relative frequencies for which the properties mentioned are easily confirmed:

1*) Relative frequency m/n cannot be negative since $m \geq 0$ and $n > 0$.

2*) By its definition, a certain event takes place in each trial and its relative frequency is therefore $n/n = 1$.

3*) If events A and B are incompatible, the event $(A + B)$ occurs when at least one of them appears, see the example above. The relative frequency of $(A + B)$ is therefore equal to the sum of the relative frequencies of A and B.

Issuing from the considerations above, we admit the three described properties of probability for any system of random events. It is useful to note that by issuing from these and some other properties we can construct axiomatically the entire probability theory. Such a strict construction is due to Kolmogorov.

1.3.1. *Remark about the subject of probability theory*. This theory only studies the numerical relations between the probabilities of various random events rather than their physical essence. The main properties of probabilities and the derived rules of calculation are here important. Indeed, the following formulation of a problem is typical for the theory and its applications.

Given, some set of simple random events whose probabilities are known. Required are the probabilities of other random events conclusively connected with the given events[3]. Thus, the occurrence of heads in each coin toss is assumed to be 1/2; determine the probability that heads appears not less than 50 times in a hundred tosses. Such problems are solved by definite rules of calculating probabilities; one of them, the addition rule, was established above.

How are the probabilities of the initial set of random events determined? For applications, it is of no consequence. It is only important that, if the relative frequencies of the initial events in a large number of trials were close to their probabilities, the same happens with the frequencies of the complex event whose probability was calculated according to the adopted rules. Those rules obey this main requirement.

1.3.2. *Corollaries of the main properties of probability*

Corollary 1. If random events $A_1, A_2, \ldots, A_n$ are pairwise incompatible,

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = P(A_1) + P(A_2) + \ldots + P(A_n). \quad (1.5)$$

Corollary 2. If such events form a complete group the sum of their probabilities is unity. Indeed, for such a group the event (A_1 or A_2 or … or A_n) is certain and

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = 1.$$

Apply formula (1.5) to the left side of this equality, then

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1. \quad (1.6)$$

Of special interest is the particular case in which the complete group only consists of two incompatible events; the occurrence of one of them is tantamount to the non-occurrence of the other. Such random events are called *mutually contrary* and denoted A and $\bar{A}$ (non-A). The sum of the probabilities of mutually contrary events is unity:

$$P(A) + P(\bar{A}) = 1. \quad (1.7)$$

1.3.3. *Impossible events.* An event is impossible if it cannot occur no matter how long we repeat the trials. Thus, it is impossible to extract a white ball from an urn that does not contain any such balls at all. An impossible event can be considered contrary to any certain event and its probability is therefore 0. This agrees with the fact that the frequency of an impossible event is also 0.

According to the classical definition of probability, the probability is zero then and only then, when the event is impossible ($M = 0$). When studying continuous random variables, we will see that a zero probability of a random event does not yet lead to its impossibility.

1.4. Products of random events. Independent events

A product of random events A and B is a random event made up of the occurrence of both A and B and is denoted by (A and B).

Example. A number is randomly chosen from the first hundred of natural numbers; events A and B are the divisibility of the selected number by 3 and 4. Then event (A and B) is the divisibility by both 3 and 4, that is, by 12. It is easily shown[4] that

$$P(A) = 33/100, P(B) = 25/100, P(A \text{ and } B) = 8/100$$

since 33 numbers are divisible by 3; 25, by 4; and 8, by 12.

The simplest relation between the probabilities of random events A and B and the probability of (A and B) takes place when these initial events are *independent*. We will first explain this notion by the pattern of random sampling. Suppose that a ball is blindly extracted from each of two urns. Events A and B will be the occurrences of a white ball from these urns respectively. These events are in essence independent since the colour of the ball extracted from one urn cannot influence the colour of the other ball. Calculate now the probability of the product of event (A and B), of both balls being white. If the urns contain N_1 and N_2 balls, M_1 and M_2 of them white, then

$P(A) = M_1/N_1$, $P(B) = M_2/N_2$.

Each of the N_1 outcomes concerning the first urn can be linked with each of the N_2 outcomes concerning the second urn so that the total number of outcomes is N_1N_2; and white balls only appear in M_1M_2 cases, so that the probability sought is

$$P(A \text{ and } B) = M_1M_2/N_1N_2 = P(A)P(B). \qquad (1.8)$$

This formula expresses the multiplication rule for independent random events. Recall that that formula is only derived for a particular case and the notion itself of independence ought to be defined. This can be done by issuing from formula (1.8) whose simplicity ensures its importance for calculations.

Definition. Two random events A and B are independent if the multiplication rule is represented for them by formula (1.8), *i. e., if the probability of their product is equal to the product of their probabilities.*

Note that independence of random events A and B leads to the pairwise independence of events $\overline{A}$ and B; A and $\overline{B}$; and $\overline{A}$ and $\overline{B}$. This statement can be easily proved in a formal way but we leave this task for the readers; see Exercise 5 in §1.6.

The definition of independence for two random events can be extended: *Events* $A_1, A_2, \ldots, A_n$ *are independent in total if the probabilities of the products of any* 2, 3, … *n of them are equal to the products of the respective probabilities.*

Thus, three events A, B and C are independent in total if

$$P(A \text{ and } B) = P(A)P(B);\ P(A \text{ and } C) = P(A)P(C);\ P(B \text{ and } C) = P(B)P(C);$$
$$P(A \text{ and } B \text{ and } C) = P(A)P(B)\, P(C). \qquad (1.9)$$

Random events can be independent pairwise but not in total. Indeed, suppose that a ball is blindly extracted from an urn having 4 balls numbered 1, 2, 3 and 123; that events A, B and C are the appearances of the numbers 1, 2 and 3 respectively on the extracted ball. These events are pairwise independent since[5]

$$P(A) = P(B) = P(C) = 2/4 = 1/2,$$
$$P(A \text{ and } B) = P(B \text{ and } C) = P(C \text{ and } A) = 1/4\ (= 1/2\cdot 1/2).$$

They are not, however, independent in total since

$$P(A \text{ and } B \text{ and } C) = 1/4 \neq 1/2\cdot 1/2\cdot 1/2$$

[here, 1/4 is the probability of extracting a ball numbered 123].

Note also that equality (1.9) all by itself does not ensure the independence of A, B and C in total. Indeed, suppose that an urn contains 8 balls numbered 1, 2, 3, 12, 13, 20, 30 and 123 and that events A, B and C are the same as above. Then

$P(A) = P(B) = P(C) = 4/8 = 1/2$;
$P(A \text{ and } B \text{ and } C) = 1/8 = 1/2(1/2\cdot 1/2)$

and even $P(A \text{ and } B) = P(A \text{ and } C) = 2/8 = 1/2\cdot 1/2$ but
$P(B \text{ and } C) = 1/8 \neq P(B)P(C)$.

1.4.1. *Generalized addition rule*. If random events A and B are independent,

$$P(A \text{ or } B) = P(A) + P(B) - P(A)P(B). \qquad (1.10)$$

Proof. Note first of all that events (A or B) and $(\bar{A} \text{ and } \bar{B})$ are contrary: the occurrence of at least one of the two events, A and B, means that the respective contrary event cannot happen, and neither the product of the contrary events, $\bar{A}$ and $\bar{B}$. By formula (1.7) and the multiplication rule we have

$$P(A \text{ or } B) = 1 - P(\bar{A} \text{ and } \bar{B}) = 1 - P(\bar{A})P(\bar{B}) =$$
$$1 - [1 - P(A)][1 - P(B)] = P(A) + P(B) - P(A)P(B), \text{ QED.}$$

I adduce now without proof the general addition rule for not necessarily independent events A and B:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \qquad (1.11)$$

Exercise 1. Prove formula (1.11) by issuing from the classical definition of probability.

Exercise 2. Geometrically interpret formula (1.11) by considering the throw of a point on a unit square as a trial and assuming that the probability of the fall of the point on some figure situated within the unit square is equal to the area of the figure.

Example 1. Two shots are independently firing at the same target. The probabilities of hitting it are $P(A) = 0.9$ and $P(B) = 0.8$ respectively. Required is the probability of at least one hit. By formula (1.10)

$$P(A \text{ or } B) = 0.9 + 0.8 - 0.9\cdot 0.8 = 0.98.$$

Example 2. Several (n) shots are firing at the same target; the probabilities of a hit are identical and equal to p. How many shots are required to hit the target with probability not lower than P?

The probability of missing the target by a shot is $1 - p$, and of each of them missing it is $(1 - p)^n$. The contrary event has therefore probability $1 - (1 - p)^n$ and the required condition is

$1 - (1 - p)^n \geq P$; therefore, $n \geq \lg(1 - P)/\lg(1 - p)$.

1.5. Conditional probabilities. General multiplication rule. The formula of total probability

We generalize the multiplication rule (1.8) on dependent random events. At first, just like in § 1.4, we consider the pattern of random sampling. The trial consists of blindly extracting a ball from an urn containing N balls identical to the touch but differing in colour and whether marked or not. We denote

K, the number of coloured balls; $(N - K)$ balls are white
L, the number of marked balls; $(N - L)$ balls are unmarked
M, the number of coloured marked balls

The events A and B are the appearances of a coloured and a marked ball and event (A and B) is then the appearance of a marked coloured ball. The probabilities of these events are

$P(A) = K/N$, $P(B) = L/N$, $P(A \text{ and } B) = M/N$.

Analogous to formula (1.8) we connect the probabilities of (A and B) and A:

$$M/N = (K/N)(M/K). \qquad (1.12)$$

The ratio M/K of the number of marked coloured balls to the number of all coloured balls is also in essence a probability, *a conditional probability of event B if event A had occurred*, and it is denoted by $P(B/A)$:

$P(B/A) = M/K$.

Now we may write (1.12) as

$$P(A \text{ and } B) = P(A)P(B/A). \qquad (1.13)$$

This relation expresses the general multiplication rule: *The probability of the product of two random variables is the product of the probability of one of them by the conditional probability of the other*.

Formula (1.13) is derived for the classical pattern. Now we turn to the general case of any random variables A and B. That formula will determine the conditional probability, the probability of event B after the occurrence of event A:

$$P(B/A) = P(A \text{ and } B)/P(A) \text{ if } P(A) \neq 0. \qquad (1.14)$$

Similarly

$$P(A/B) = P(A \text{ and } B)/P(B) \text{ if } P(B) \neq 0. \qquad (1.15)$$

It is not difficult to verify that conditional probabilities possess all the main properties of probabilities.

Formula (1.13) can be generalized on a larger number of random events. For three of them

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

The notion of conditional probability allows us to interpret independence of random events anew. If random events A and B are independent, then, by formulas (1.8, 1.14 and 1.15),

$$P(B/A) = P(A)P(B)/P(A) = P(B),\ P(A/B) = P(A)P(B)/P(B) = P(A)$$

which means that the conditional and unconditional probabilities of each of these events coincide. It is also obvious that, inversely, if $P(B/A) = P(B)$, formula (1.13) is transformed into (1.8).

The independence of random events A and B thus means that the probability of B (or A) does not change when an additional condition of the occurrence of A (or B) is introduced. This interpretation directly leads, for example, to the independence of a certain event and any random event A.

To recall, the probability of random event B is invariably connected with a definite set of conditions. Add to it the occurrence of some other event A, and the probability of B can change.

In the example above, the independence of random events A and B was reduced to the equality

$$M/K = L/N,$$

i. e., to the condition that the number of marked and coloured balls is to the number of coloured balls as the number of marked balls to the total number of balls in the urn.

1.5.1. *The formula of total probability*

Theorem. If random events $H_1, H_2, \dots, H_n$ are pairwise incompatible and event A can only occur together with one of them, then

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2) + \dots + P(H_n)P(A/H_n). \quad (1.16)$$

Proof. Event A is tantamount to the product of events

$$[(H_1 \text{ or } H_2 \text{ or } \dots \text{ or } H_n) \text{ and } A].$$

However, this product occurs then and only then when one of the products (H_1 and A) or (H_2 and A) … or (H_n and A) takes place. By the addition rule

$$P(A) = P[(H_1 \text{ or } H_2 \text{ or } \dots \text{ or } H_n) \text{ and } A] =$$
$$P(H_1 \text{ and } A) + P(H_2 \text{ and } A) + \dots + P(H_n \text{ and } A). \quad (1.17)$$

It only remains to apply the general multiplication rule:

$P(H_1 \text{ and } A) = P(H_1)P(A/H_1); \ldots$

In particular, the formula

$$P(B) = P(B)P(B/A) + P(\overline{A})P(B/\overline{A}) \qquad (1.18)$$

invariably takes place since the contrary events A and $\overline{A}$ are incompatible and make up a complete group.

Example 1. An urn contains N balls, M of them white. Two balls are extracted one after the other. Events A and B are the respective extractions of a white ball. Obviously

$$P(A) = M/N,\ P(\overline{A}) = (N - M)/N,$$
$$P(B/A) = (M - 1)/(N - 1),\ P(B/\overline{A}) = M/(N - 1).$$

By formula (1.18)

$$P(B) = \frac{M}{N}\frac{M-1}{N-1} + \frac{N-M}{N}\frac{M}{N-1} = \frac{M}{N}.$$

The probabilities of the appearance of a white ball at each extraction are identical.

Example 2. A ball is thrown on a board and can stop at points (x_k, y_l), $k = 1, 2, \ldots, n$; $l = 1, 2, \ldots, m$. The probabilities of the ball occurring in those positions are p_{kl}:

$$\begin{aligned} &p_{11}, p_{21}, \ldots, p_{n1}, \text{ if } y = y_1 \\ &p_{12}, p_{22}, \ldots, p_{n2}, \text{ if } y = y_2, \ldots, \\ &p_{1m}, p_{2m}, \ldots, p_{nm}, \text{ if } y = y_m \end{aligned} \qquad (1.19)$$

Required is the probability of the ball falling in column x_k. Abscissa x_k can only be connected with one of the ordinates $y_1, y_2, \ldots, y_m$ so that by formula (1.17) we have

$$P_k = p_{k1} + p_{k2} + \ldots + p_{km},\ k = 1, 2, \ldots, n. \qquad (1.20)$$

1.6. Exercises

1) Three cards are extracted from a deck (of 52 cards). Required is the probability that there will be at least one ace. *Answer*: $p = 1201/5525 = 0.217$.

2) An operator services three independently functioning lathes. During an hour they need the operator's attention with probabilities 0.1, 0.2 and 0.3. Required is the probability that during an hour the operator will have to attend to at least one lathe. *Answer*: $p = 0.994$.

3) An urn contains 5 white and 20 black balls. They are extracted one by one until a white ball appears. Required is the probability that there will be 3 extractions (that 2 black balls will appear before a white ball). *Answer*: $p = 19/138 = 0.138$.

4) Machine parts of the same type are produced by two lathes. The probabilities of the appearance of substandard parts are 0.03 and 0.02

respectively. The finished parts are stored together, twice more of them from the first lathe than from the second. Required is the probability that a randomly chosen part will not be substandard. *Answer*. $p = 292/300 = 0.973$.

5) By issuing from formula (1.18) prove that the independence of random events *A* and *B* leads to independence of $\overline{A}$ and *B*.

Chapter 2. Random Variables and Distribution of Probabilities

2.1. Discrete random variables

Here, we study variables whose values depend on chance (the number of points achieved when casting a die; the number of calls entering a telephone exchange during given time etc).

Definition. Magnitude ξ *is called a discrete random variable if all of its values form a finite or infinite number sequence* $x_1, x_2, \ldots, x_k, \ldots$ *and the appearance of each of them* ($\xi = x_k$) *is a random event having a definite probability.*

We denote the probability of $\xi = x_k$ by p_k, a function of x_k. This function is called *the law of distribution of probabilities of* ξ. *Any rule that allows us to determine the probabilities of all the possible values of* ξ *determines the distribution of its probabilities.*

The two rows, $x_1, x_2, \ldots, x_k, \ldots$ and $p_1, p_2, \ldots, p_k, \ldots$, form a table of that distribution. If ξ can only take a finite number of differing values $x_1, x_2, \ldots, x_n$, the random events $\xi = x_1$, $\xi = x_2$, …, $\xi = x_n$ form a complete group of incompatible events and the sum of their probabilities is unity

$$p_1 + p_2 + \ldots + p_n = 1. \qquad (2.1)$$

For the case of infinitely many values of ξ the series $p_1 + p_2 + \ldots + p_n + \ldots$ should converge and [as before] its sum ought to be unity.

Example 1. The number of points on a die is a discrete random variable with the table of distribution:

values: 1, 2, 3, 4, 5, 6; probabilities: 1/6, 1/6, … (2.2)

For an irregular die the values are the same, but their probabilities will not be identical.

Example 2. A hunter has 3 cartridges and shoots until he hits the target (or expends all of them). The number of expended cartridges is a random variable ξ with possible values 1, 2 and 3. The probability of a hit is 0.8 and required is the distribution of probabilities. We have

$$P(\xi = 1) = 0.8;\ P(\xi = 2) = (1 - 0.8)\cdot 0.8 = 0.2\cdot 0.8 = 0.16.$$

Here, $P(\xi = 2)$ is the product of two probabilities, of missing the first time and then hitting the target. The last probability can be calculated either directly or by formula (2.1):

$$P(\xi = 3) = 0.2\cdot 0.2 = 0.04;\ \text{or, } P(\xi = 3) = 1 - P(\xi = 1) - P(\xi = 2).$$

The table of the distribution of probabilities is:

values of ξ: 1, 2, 3; probabilities: 0.8, 0.16, 0.04 (2.3)

Example 3. Shots fire at a target until hitting it. The probability of a hit is p, and the number of attempts is a random variable ξ with an infinite table of the distribution of probabilities

values: 0, 1, 2, n, …
probabilities p: $(1 - p)p$, $(1 - p)^2p$, …, $(1 - p)^{n-1}p$, … (2.4)

These probabilities form an infinitely decreasing geometrical progression with ratio $(1 - p)$. It converges and its sum is $p/[1 - (1 - p)] = 1$.

Example 4. In some physical and technical problems (when considering the number of calls entering an automatic telephone exchange or of electrons flying out from an incandescent cathode, in both cases during a certain period of time, etc), there appear random variables obeying the Poisson law of distribution

0, 1, 2, …, m, …
$e^{-a}(1, a, a^2/2, …, a^m/m!, …)$ (2.5)

Here, a is some positive number, see below. The series of probabilities converges and its sum is unity:

$$e^{-a}(1+ a + a^2/2 + … + a^m/m! + …) = e^{-a}\cdot e^a = 1.$$

2.1.1. *Linear operations on random variables*. We have in mind multiplication of a random variable by a number and addition of random variables.

The product $C\xi$ of a discrete random variable ξ by number C is a random variable with the distribution of probabilities

values: Cx_1, Cx_2, …; probabilities: p_1, p_2, … (2.6)

All the values of ξ are multiplied by C, but their probabilities remain unchanged.

Somewhat more complicated is the distribution of probabilities of a sum of two discrete random variables with values of ξ being x_1, x_2, … and of η being y_1, y_2, … and probabilities p_1, p_2, … and q_1, q_2, … Denote the probability of the product of the random events $\xi = x_k$ and $\eta = y_l$ by p_{kl}. However, the probability of, say, the event $\xi + \eta = x_1 + y_1$, can exceed p_{11} if among the sums $x_k + y_l$ there are numbers equal to $x_1 + y_1$. Indeed, according to the addition rule, we ought to consider the probability of the event $\xi + \eta = x_1 + y_1$ equal to the sum of all those probabilities p_{kl} for whom $x_k + y_l = x_1 + y_1$.

Therefore, the values of the sum $\xi + \eta$ are the sums of all the possible values of ξ and η, and the probability of each is the sum of the probabilities of the products of $\xi = x_k$ and $\eta = y_l$ for which their sum takes that value. […]

Example. The trial consists of casting at once two regular dice. Denote by ξ and η the numbers of points appearing on them, and $\xi + \eta$ will be the sum of these points. These ξ and η are random variables and have the same table of the distribution of probabilities (2.2). Required is the distribution of their sum, $\xi + \eta$. The outcomes of the casting are independent, so that the probability of each product will be 1/36. An auxiliary table will be

sums: $1 + 1,\ 1 + 2,\ 2 + 1,\ 2 + 2,\ 3 + 1, \dots, 6 + 6$
probabilities: identically 1/36

After combining the equal sums we get the final table of the probabilities of the values of $(\xi + \eta)$:

2	3	…	6	7	8	…	11	12	
1/36	2/36	…	5/36	6/36	5/36	…	2/36	1/36	

Note a singularity: when comparing this table with that for the distribution of 2ξ

2, 4, 6, 8, 10, 12 and identical probabilities 1/6

we conclude that the addition of random variables with identical distributions of probabilities is not in general reduced to multiplying one of them by an integer.

Addition of random variables or their multiplication by a number does not change the known properties of addition and multiplication of numbers. In particular,

$$\xi + \eta = \eta + \xi;\ (\xi + \eta) + \zeta = \xi + (\eta + \zeta);\ C(\xi + \eta) = C\xi + C\eta.$$

2.1.2. *Independence of random variables*. Discrete random variables ξ and η are *independent*, if the random events $\xi = x_k$ and $\eta = y_l$ are independent for all values of k and l. In other words, if

$$p_{kl} = p_k q_l,\ k = 1, 2, \dots\ l = 1, 2, \dots \qquad (2.7)$$

For example, when casting two dice, the numbers of points on either are independent random variables. This has indeed simplified the calculation of their sum above.

Random variables $\xi_1, \xi_2, \dots, \xi_n$ are mutually independent if all the random events $\xi_1 = x_1$, $\xi_2 = x_2, \dots, \xi_n = x_n$ are independent in total. Magnitudes $\{x\}$ are the values of the given random variables ξ_i.

So if random variables $\xi_1, \xi_2, \dots, \xi_n$ are mutually independent and the distribution of their probabilities is given, the distribution of the probabilities of any of their linear functions with constant coefficients

$$C_1\xi_1 + C_2\xi_2 + \dots + C_n\xi_n$$

can be easily determined. This circumstance is often made use of by representing a studied random variable as a linear function of independent random variables with a known distribution of probabilities.

2.2. The distribution of probabilities of the relative frequency of random events

We consider the frequency w_n of a random event A after n trials. Suppose that the occurrence of A in each trial has the same probability p which does not depend on the results of other trials. Such a repetition of trials can be imagined as extractions with replacement of balls of two different kinds from an urn and it is called *the sequence of independent trials according to the Bernoulli pattern* (*or to the pattern of the replaced balls*).

In n trials the event A can occur 0, 1, 2, …, n times and therefore w_n is a discrete random variable with possible values 0, $1/n$, $2/n$, …, 1. Required is the distribution of that frequency. Represent w_n as a linear combination of simpler random variables by introducing the so-called indicator random variables λ_k, the number of the occurrences of event A in the k-th trial. It only takes 2 values, 1, if the event occurs, and 0 otherwise. The probability of event A in each trial is p, so these indicator variables have an identical distribution of probabilities,

values: 1 and 0; probabilities: p and q, $q = 1 - p$. (2.8)

According to the adopted condition, all the variables $\lambda_1, \lambda_2, \dots, \lambda_n$ are independent in total. Consider now their sum

$$\mu_n = \lambda_1 + \lambda_2 + \dots + \lambda_n. \qquad (2.9)$$

The terms are unities or zeros and there are exactly as many unities as many times (μ_n) the event A happens in n trials. The ratio μ_n/n is the relative frequency w_n:

$$w_n = \mu_n/n = (1/n)(\lambda_1 + \lambda_2 + \dots + \lambda_n). \qquad (2.10)$$

This representation of w_n by a linear combination of mutually independent random variables $\lambda_1, \lambda_2, \dots, \lambda_n$ with a known distribution of probabilities (2.8) allows us to determine the distribution of the probabilities of w_n. Sum up these random variables consecutively; by formula (2.7) we have $\lambda_1 + \lambda_2 = w_2$ with the distribution of probabilities being:

values: 2, 1, 0; probabilities: pp, $pq + qp$, qq, or p^2, $2pq$, q^2

In the same way we determine the law of the distribution of w_3 and note that the calculated probabilities coincide with the corresponding terms of the expansion of binomials $(p + q)^2$ and $(p + q)^3$. Since $p + q = 1$, it is seen at once that the sums of the probabilities in the tables of the distribution of probabilities equal 1. It is possible to prove the following general statement by mathematical induction: the probability that μ_n takes some value m is equal to the term which includes p^m in the expansion of the binomial $(p + q)^n$ in powers of p:

$$P(\mu_n = m) = C_n^m p^m q^{n-m}. \qquad (2.11)$$

It is possible to derive formula (2.11) by complete induction without issuing from (2.9). Moreover, the probability that A occurs in the first m trials and will not arrive in the other $(n - m)$ trials can be calculated by the multiplication rule for independent events. We get

$$p^m q^{n-m}. \qquad (2.12)$$

This probability does not depend on the choice of the successful trials which can be made in C_n^m different ways. The probability of the event $\mu_n = m$ is therefore equal to (2.12) multiplied by C_n^m, QED.

Formula (2.11) provides the probability that event A will occur exactly m times in n trials. Therefore we have the following tables of the distribution of probabilities for random variables μ_n and w_n:

values: n, $(n - 1)$, …, m, …, 1, 0; [binomial probabilities] (2.13)

values: 1, $(n - 1)/n$, …, $1/n$, …, 0; [the same probabilities] (2.14)

The distribution determined by table (2.13) is called *binomial.*

Example. The quality of a large batch of machine parts is checked by a sample of 10. It is known that there are 25% substandard parts in the entire batch [which should therefore be thrown away in its entirety]. It is required to determine the probability that more than 5 parts in the sample are substandard.

The selection of each part for the sample is a trial, and its being substandard is a random event A. Its probability is obviously $p = 0.25$, and we have to determine the probability of $\mu_{10} > 5$. By the addition rule

$$P(\mu_{10} > 5) = P(\mu_{10} = 6) + P(\mu_{10} = 7) + \ldots + P(\mu_{10} = 10).$$

These probabilities can be calculated by the binomial formula (2.11) for $p = 0.25$, $q = 0.75$ and $n = 10$. [The author provides a table for $\mu_{10} = 0(1)10$.] So $P(\mu_{10} > 5) \approx 0.020$, a rather low probability.

[The author notes that the solution is only rigorous if the sample is made with replacement but that this restriction becomes the less important the larger is the size of the sample.]

2.3. Continuous random variables

The theory of probability often has to consider random variables whose possible values completely fill up some interval, with continuous random variables, see for example the Introduction. The law of the distribution of probabilities of such a variable ξ should allow us to determine the probability of its value to be in any interval (x_1, x_2), $P(x_1 < \xi < x_2)$ [contained within the initial interval].

Example: The uniform distribution of probabilities. In the simplest case all the possible values of such random variable ξ fill up some finite interval (α_1, α_2) and the probability $P(x_1 < \xi < x_2)$ for any interval (x_1, x_2) situated within (α_1, α_2) is proportional to its length:

$$P(x_1 < \xi < x_2) = \lambda(x_2 - x_1),\ \alpha_1 \le x_1 < x_2 \le \alpha_2. \qquad (2.15)$$

The coefficient λ should ensure the second main property of probabilities; the first property is secured by taking $\lambda > 0$, and the third follows from the addition of the lengths of the intervals necessary when they are combined.

Since all the possible values of ξ are situated within (α_1, α_2), this random variable is certainly there also:

$$P(\alpha_1 < \xi < \alpha_2) = \lambda(\alpha_2 - \alpha_1) = 1, \text{ and } \lambda = 1/(\alpha_2 - \alpha_1).$$

The random variable ξ is uniformly distributed on interval (α_1, α_2) *if the distribution of its probability is described by formula* (2.15).

2.3.1. *Density of the distribution of probabilities*. For a random variable ξ uniformly distributed on interval (α_1, α_2) the ratio

$$P(x_1 < \xi < x_2)/(x_2 - x_1) \qquad (2.16)$$

is constant and equal to $\lambda = 1/(\alpha_2 - \alpha_1)$. This ratio is called the density of the distribution of probabilities for a uniformly distributed random variable ξ.

That ratio is not in general constant. We have to introduce the notion of density in a given point, just like it is done in physics when considering the distribution of the mass of a body.

The density of the distribution of probabilities of random variable ξ in point x is the limit as $\Delta x \to 0$

$$\lim \frac{P(x < \xi < x + \Delta x)}{\Delta x} = \varphi(x). \qquad (2.17)$$

We only consider random variables for which that limit exists in each point x. For them, the probability of the value of ξ to be within the interval $(x, x + \Delta x)$ is

$$P(x < \xi < x + \Delta x) \approx \varphi(x)dx$$

to within small magnitudes of higher orders. This main part of the probability is called its differential:

$$dP_x = \varphi(x)dx. \qquad (2.18)$$

Given this dP_x, we can by integrating determine the probability of the value of ξ to be within any interval (x_1, x_2):

$$P(x_1 < \xi < x_2) = \int_{x_1}^{x_2} \varphi(x)dx. \qquad (2.19)$$

Thus, *for determining the law of the distribution of a continuous random variable it is sufficient to provide the density of the distribution of its probabilities, i. e., the function* $\varphi(x)$.

Strictly speaking, a continuous random variable ξ is indeed characterized by representing the probability $P(x_1 < \xi < x_2)$ as the integral (2.19) of some function $\varphi(x)$.

In any calculations involving a continuous random variable the differential $\varphi(x)dx$ plays the same role as the probabilities p_k when dealing with discrete random variables. In many formulas it is sufficient to replace p_k by $\varphi(x)dx$ and substitute the sum for the corresponding integral.

Remark. The occurrence of a value of a continuous random variable in an isolated point is of no consequence, *the probability of that event is zer*o. Significant is only its taking place on some interval. The probability of the random variable to be on a short interval is approximately proportional to the length of that interval. In other words, if the continuous random variable taking a definite value is considered as a random event, then the probability of that event ought to be zero (although it cannot be thought as impossible). The statement above does not lead to confusion since the value of any physical magnitude can only be measured with some precision; absolutely precise values of such magnitudes are only mathematical abstractions.

2.3.2. *The main properties of the density of the distribution.*

1) The density $\varphi(x)$ is non-negative for all values of x. This directly follows from definition (2.17) in which $\Delta x > 0$ and $P(x < \xi < x+ \Delta x) \geq 0$.

2) The integral of density $\varphi(x)$ over the domain of random variable ξ is unity which follows from the meaning of that integral: it expresses the probability of a certain event, of ξ taking some of its values. This property is expressed as

$$\int_{\alpha_1}^{\alpha_2} \varphi(x)dx = 1 \text{ or } \int_{-\infty}^{\infty} \varphi(x)dx = 1, \text{ or, in general, } \int \varphi(x)dx = 1 \quad (2.20)$$

when assuming that the last integral is taken over the domain of ξ.

Each non-negative function $\varphi(x)$ satisfying condition (2.20) can be a density of some random variable.

The curve of the distribution of probabilities is the graph of the density $y = \varphi(x)$. It can serve for graphically reckoning probabilities since the probability $P(x_1 < \xi < x_2)$ is expressed by the same integral as is the corresponding area *under* $\varphi(x)$ if only the complete area under it is unity. The probability $P(x_1 < \xi < x_2)$ *is equal to the ratio of those areas*.

2.3.3. *Examples of continuous distributions of probabilities*.

1) The simplest normal distribution. A continuous random variable ξ_0 obeys that law if

$$\varphi_0(x) = C\exp(-x^2/2),\ C = 1/\sqrt{2\pi}. \qquad (2.21)$$

The value of C thus ensures the condition (2.20) and the curve is symmetric with respect to the y-axis. It takes its maximal value $1/\sqrt{2\pi} \approx 0.4$ at $x = 0$ and has 2 points of inflexion, $x = \pm 1$. At $x \to \pm\infty$ the curve very rapidly tends to the x-axis; thus, $\varphi_0(3) = 0.0044$ and $\varphi_0(4) = 0.00013$.

The normal distribution is very important in many applications, in particular for treating observations (Chapters 5 and 6). The integral of $\varphi_0(x)$ cannot be expressed in a finite way by elementary functions and very detailed and sufficiently precise tables of the integral of probability

$$\Phi(t) = \frac{2}{\sqrt{2\pi}}\int_0^t \exp(-x^2/2)dx \qquad (2.22)$$

have been therefore compiled.

This function is odd: $\Phi(-t) = -\Phi(t)$ so that the tables only provide the values of $\Phi(t)$ for $t > 0$. When t changes from 0 to ∞, $\Phi(t)$ very rapidly increases from 0 to 1. Thus, $\Phi(3) = 0.9973$, $\Phi(4) = 0.999937$.

It is possible to apply function $\Phi(t)$ for calculating probabilities:

$$P(x_1 < \xi < x_2) = \frac{1}{\sqrt{2\pi}}[\int_0^{x_2} \exp(-x^2/2)dx - \int_0^{x_1} \exp(-x^2/2)dx] =$$
$$\frac{1}{2}[\Phi(x_2) - \Phi(x_1)]. \qquad (2.23)$$

For a symmetric interval $(-t, t)$

$$P(-t < \xi < t) = \frac{1}{2}[\Phi(t) - \Phi(-t)] = \Phi(t). \qquad (2.24)$$

2) The general normal distribution of probabilities has density

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-a)^2}{2\sigma^2}]dx,\ \sigma > 0. \qquad (2.25)$$

For $a = 0$ and $\sigma = 1$ this density becomes $\varphi_0(x)$. With an increasing σ the curves of the normal distribution become more sloping whereas the change of a leads to a shifting of the curve along the x-axis.

3) An example of an asymmetric distribution of probabilities with density

$$\varphi(x) = 0 \text{ at } x \le 0; \text{ and, at } x > 0 \text{ and } \alpha, \beta > 0,\ C_1 x^{\alpha-1}e^{-\beta x}. \quad (2.26)$$

C_1 is chosen to ensure the fulfilment of condition (2.20):

$$C_1 = \beta^\alpha/\Gamma(\alpha),\ \Gamma(\alpha) = \int_0^\alpha x^{\alpha-1}e^{-x}dx\,.$$

Here Γ(α) is the Euler gamma-function. Distribution (2.26) belongs to the so-called Pearsonian curves and occurs in many problems connected with hydroelectricity. Calculations involving this distribution are again possible by means of special tables.

2.3.4. *The distribution function.* The distribution function of the probabilities of random variable ξ is the probability that ξ takes a value smaller than x:

$$F(x) = P(\xi < x).$$

For a discrete random variable it is equal to the sum of those of its values which are smaller than x:

$$F(x) = \sum p_k,\ x_k < x.$$

Thus, for a random variable with a table of distribution (2.3)

$$F(x) = 0,\ x \le 1;\ 0.8,\ 1 < x \le 2;\ 0.96,\ 2 < x \le 3;\ 1,\ x > 3.$$

According to formula (2.19), the distribution function is the integral of the density

$$F(x) = \int_{-\infty}^{\infty} \varphi(t)dt.$$

For example, in case of the simplest normal distribution (2.21) it is

$$F(x) = \int_{-\infty}^{x} \varphi_0(t)dt = \frac{1}{2}[\Phi(x) - \Phi(-\infty)] = \frac{1}{2}\Phi(x) + \frac{1}{2}.$$

The main properties of probability and the definition of $F(x)$ require that that function is increasing and takes values from 0 to 1. Its graph is called the integral curve of the distribution of probabilities. Then, for $x_1 < x_2$

$$P(\xi < x_2) = P(\xi < x_1) + P(x_1 \le \xi < x_2),\ P(x_1 \le \xi < x_2) = F(x_2) - F(x_1).$$

The distribution function can be applied for describing the law of distribution both for discrete and continuous (and more complicated) random variables. However, in general such applications require the use of a special mathematical tool, the Stieltjes integral.

2.4. Functions of random variables

Suppose that $f(x)$ is a one-valued function determined for all possible values x of random variable ξ. Function $f(\xi)$ is understood as a random variable η that takes value $y = f(x)$ whenever ξ takes value x.

For example, if ξ is the diameter of a cylinder turned on a lathe, the area of its cross-section is random variable $\eta = (\pi/4)\xi^2$.

We ought to establish the connection between the laws of the distribution of probabilities of ξ and η and we begin with functions of a discrete random variable ξ having values $x_1, x_2, \ldots$ and probabilities $p_1, p_2, \ldots$ If ξ takes value x_k, then $\eta = f(x_k)$. However, the probability of, say, $\eta = f(x_1)$ can be higher than p_1 if among the values of $f(x_k)$ there are numbers equal to $f(x_1)$. Indeed, by the addition rule we must assume that that probability is equal to the sum of all the probabilities p_k for which $f(x_k) = f(x_1)$.

For constituting a table of distribution for $f(\xi)$ we usually begin with an auxiliary table

$$f(x_1), f(x_2), \ldots; \text{ probabilities } p_1, p_2, \ldots, \qquad (2.27)$$

then sum up the probabilities of identical values of $f(x_k)$. However, if all the values of $f(x_k)$ are different, the table (2.27) will be final for the distribution of function $f(\xi)$, see for example (2.6).

Example 1. Consider powers λ^n, $n = 1, 2, 3, \ldots$ of the indicator variable λ with distribution of probabilities (2.8). All these powers have the same distribution as λ itself since $1^n = 1$, $0^n = 0$.

Example 2. Consider the function $\sin[(\pi/2)\xi]$ of random variable ξ with values $1, 2, \ldots, n, \ldots$ and probabilities $1/2, 1/2^2, \ldots, 1/2^n, \ldots$ Since

$$\sin[(\pi n/2)] = 0 \text{ if } n \text{ is even; } 1, \text{ if } n = 4k+1; -1 \text{ if } n = 4k+3,$$

$\sin[(\pi/2)\xi]$ takes values $0, 1, -1$ with probabilities p_0, p_1, p_{-1} $[k = 0, 1, 2, \ldots]$,

$$p_0 = 1/2^2 + 1/2^4 + 1/2^6 + \ldots = 1/3; \; p_1 = 1/2 + 1/2^5 + 1/2^9 + \ldots = 8/15;$$
$$p_{-1} = 1/2^3 + 1/2^7 + 1/2^{11} + \ldots = 2/15.$$

Example 3. Random variable ξ takes values $-b$ and b with probabilities p and $1-p$. Consider its square, ξ^2; it takes a single value b^2 with probability 1 and can be thought as being non-random.

2.4.1. *Functions of continuous random variables*. Given, function $f(x)$, continuous as is its first derivative over the interval of all possible values of x of a random variable ξ. Establish the dependence between the densities of the probabilities of $\varphi(x)$ and $\psi(y)$ of random variables ξ and $\eta = f(\xi)$.

This aim is attained in the simplest way when $f(x)$ *strictly increases* so that each interval (x_1, x_2) is in a one-to-one correspondence with interval (y_1, y_2). The probabilities of ξ and η to be in those intervals are therefore identical. For short intervals $(x, x+\Delta x)$ and $(y, y+\Delta y)$ this leads to the equality of the differentials of the probabilities

$$\varphi(x)dx = \psi(y)dy. \qquad (2.28)$$

Therefore

$$\psi(y) = \varphi(x)dx/dy = \varphi[g(y)]g'(y) \qquad (2.29)$$

where $x = g(y)$ is the function inverse to $y = f(x)$.

If $y = f(x)$ *strictly decreases*, a negative dy corresponds to a positive dx and therefore dy in formula (2.28) should be replaced by $-dy = |dy|$. In general, we thus obtain

$$\psi(y) = \varphi(x)|dx/dy| = \varphi[g(y)]|g'(y)|. \qquad (2.30)$$

Example. A linear function $\eta = a + b\xi$. We have

$y = f(x) = a + bx$, $x = g(y) = (y - a)/b$, $g'(y) = 1/b$ and

$$\psi(y) = (1/|b|)\varphi[(y - a)/b]. \qquad (2.31)$$

If random variable ξ is uniformly distributed on (α_1, α_2), the random variable $\eta = a + b\xi$ will also be uniformly distributed on $(a + b\alpha_1, a + b\alpha_2)$. We leave the proof to the readers.

Suppose now that random variable ξ_0 has the simplest normal distribution with density (2.21). Then $\eta = a + b\xi_0$ will have the general normal distribution with density

$$\psi(y) = \frac{1}{|b|\sqrt{2\pi}}\exp[-\frac{(y-a)^2}{2b^2}].$$

This conclusion allows us to calculate probabilities for the general normal distribution (2.25) by formula (2.22). Indeed, suppose that random variable ξ has that distribution with density (2.25). Then random variable $\xi_0 = (\xi - a)/\sigma$ will have the simplest normal distribution (2.21) and the inequalities $x_1 < \xi < x_2$ will be tantamount to

$$\frac{x_1 - a}{\sigma} < \xi_0 < \frac{x_2 - a}{\sigma}.$$

Therefore

$$P(x_1 < \xi < x_2) = P(\frac{x_1 - a}{\sigma} < \xi_0 < \frac{x_2 - a}{\sigma}) = \frac{1}{2}[\Phi(t_2) - \Phi(t_1)]. \quad (2.32)$$

Here, $t_1 = (x_1 - a)/\sigma$, $t_2 = (x_2 - a)/\sigma$.

2.4.2. *Deriving the density of distribution for a non-monotone function*. We only consider the function $\eta = \xi^2$ with the unbounded domain $(-\infty, \infty)$ for ξ. Here, $y = f(x) = x^2 \geq 0$. The inverse function has two one-valued branches: $x = g_1(y) = \sqrt{y}$ and $x = g_2(y) = -\sqrt{y}$. When applying formula (2.30) to each of these and combining equal values of y, we will have for $y > 0$

$$\psi(y) = [\varphi(\sqrt{y}) + \varphi(-\sqrt{y})]\frac{1}{2\sqrt{y}} \text{ if } y > 0, \text{ and } \psi(y) = 0 \text{ if } y < 0.$$

2.4.3. *Notion of two-dimensional random variables and functions of two random variables*. For solving many problems as well as for studying functions of several random variables it is necessary to consider many-dimensional random variables, i. e. those whose values are distributed in two-, three- and many-dimensional spaces. An example of a two-dimensional random variable is a hit-point on a target. Denoting its coordinates by ξ and η, we get a two-dimensional random variable (ξ, η), see a similar example of a discrete two-dimensional variable in Example 2 of § 1.5.1 and the table (1.19) of the distribution of its probabilities.

I only indicate some formulas describing continuous two-dimensional random variables; appropriate formulas for discrete two-dimensional variables are similar. The value of variable (ξ, η) is point (x, y) and the distribution of the probabilities is given by the differential of probability

$$dP_{xy} = \varphi(x, y)dxdy. \qquad (2.33)$$

It provides the main part of the probability that point (ξ, η) is in the rectangle $x < \xi < x + dx$, $y < \eta < y + dy$. Function $\varphi(x, y)$ is called the two-dimensional density of distribution. The probability of point (ξ, η) to be within some region D is determined by the integral over D

$$P[(\xi, \eta) \in D] = \int\int \varphi(x, y)dxdy. \qquad (2.34)$$

Any non-negative function satisfying condition

$$\int\int \varphi(x, y)dxdy = 1$$

with the integral covering all possible values of random variable (ξ, η) can be a density.

The simplest example of a continuous two-dimensional random variable is variable (ξ, η) having a uniform distribution on some finite region D_0. For such variables the probability of their being in any region D within D_0 is proportional to the area S_D and

$dP_{xy} = \lambda dxdy$ for points within D_0 and 0 otherwise.

The coefficient λ is defined by the condition

$$P[(\xi, \eta) \in D_0] = \lambda S_{D_0} = 1.$$

And if D is within D_0, the probability of (ξ, η) to be within D is the ratio of the respective areas. The two-dimensional density is here $\varphi(x, y)$ = 1/(area of D_0).

The coordinates ξ and η of a two-dimensional continuous random variable are one-dimensional continuous random variables. Their densities $\psi_1(x)$ and $\psi_2(y)$ are connected with the two-dimensional density $\varphi(x, y)$ by formula

$$\psi_1(x) = \int \varphi(x, y)dy, \ \psi_2(y) = \int \varphi(x, y)dx. \qquad (2.35, 2.36)$$

For justifying, say, (2.35) suffice it to note that the differential of probability $\psi_1(x)dx$ can be considered as the probability of point (ξ, η) to be in an infinite band between vertical lines having abscissas x and $x + dx$. For a discrete variable with distribution (1.19) the similar formula is (1.20).

Random variables ξ and η are independent if

$dP_{xy} = \psi_1(x)dx\psi_2(y)dy$; that is, if

$$\varphi(x, y) = \psi_1(x)\psi_2(y). \qquad (2.37)$$

For a discrete variable with distribution (1.19) the similar formula is (2.7).

For function $\varsigma = f(\xi, \eta)$ of two random variables ξ and η the distribution of probabilities is

$$P(z < \varsigma < z + dz) = \iint \varphi(x, y)dxdy \qquad (2.38)$$

The integral is taken over such a region of the plane (x, y) that $z < f(x, y) < z + \Delta z$. Here, $\varphi(x, y)$ is the density of distribution of the two-dimensional random variable (ξ, η). When isolating the main part, linear with respect to Δz in the integral (2.38), we thus determine the differential dP_z and therefore the density of the distribution of function $\varsigma = f(\xi, \eta)$.

2.4.4. *The distribution of a sum of random variables*. For the sum $\varsigma = \xi + \eta$ the region covered by integral (2.38) is a band between straight lines $x + y = z$ and $x + y = z + \Delta z$. Therefore

$$P(z < \zeta < z + \Delta z) = \int_{-\infty}^{\infty} dx \int_{z-x}^{z+\Delta z-x} \varphi(x, y)dy.$$

The main part of the inner integral is approximately $\varphi(x, z - x)\Delta z$ and the differential of the probability of ς is

$$dP_z = \int_{-\infty}^{\infty} \varphi(x, z - x)dx\Delta z.$$

.

The density $\chi(z)$ of the distribution of the sum $\varsigma = \xi + \eta$ will be

$$\chi(z) = \int_{-\infty}^{\infty} \varphi(x, z - x)dx. \qquad (2.39)$$

Especially interesting is the case of independent ξ and η. Formula (2.37) allows us then to express the density of that sum by the densities of ξ and η:

$$\chi(z) = \int_{-\infty}^{\infty} \psi_1(x)\psi_2(x, z-x)dx. \qquad (2.40)$$

This integral is called the convolution of ψ_1 and ψ_2 and denoted by $\psi_1 * \psi_2$.

2.5. Exercises

1) Determine the distribution of the probability of the sum of points on three dice. Check that the outcome of 11 points is more probable than 12 points although both are realized in 6 ways:

11 points: 6, 4, 1; 6, 3, 2; 5, 5, 1; 5, 4, 2; 5, 3, 3; 4, 4, 3
12 points: 6, 5, 1; 6, 4, 2; 6, 3, 3; 5, 5, 2; 5, 4, 3; 4, 4, 4

2) An urn contains 20 black and 4 white balls. Determine the distribution of the probabilities of the number of white balls after 5 balls have been extracted. *Answer*:

values: 0, 1, 2, 3, 4; prob.: (1/1771)(646, 1615, 285, 95, 5)

3) Balls are extracted from the same urn until a black ball appears. Determine the distribution of the probabilities of the number of white balls which had appeared previously. *Answer*:

values: 0, 1, 2, 3, 4; prob.: 5/6, 10/69, 5/253, 10/5313; 1/10,626

4) Determine the [density of the] sum of two independent variables uniformly distributed on interval (– 1, 1). Answer:

$\varphi(x) = 0$ if $x < -2$ or $x > 2$; $(1/4)(x + 2)$ if $-2 < x < 0$;
$(1/4)(-x + 2)$ if $0 < x < 2$

5) A point falls on a circumference and its position along the circumference is uniformly distributed. Required is the distribution of the probabilities of the projection of that point on a diameter. *Answer*:

$\varphi(x) = 0$ if $x < -R$ or $x > R$; $\dfrac{1}{\pi\sqrt{R^2 - x^2}}$ otherwise

6) Suppose that the linear dimensions of somewhat irregular cubes are normally distributed (2.25). Determine the distribution of the probabilities of the volumes of these cubes. *Answer*: the density is

$$\psi(v) = \frac{1}{3\sqrt[3]{v^2}\sigma\sqrt{2\pi}}\exp[-\frac{(\sqrt[3]{v} - a)^2}{2\sigma^2}].$$

7) Prove that the most probable value of frequency μ_n is integer m_0 such that $np + p - 1 \le m_0 \le np + p$; if np is an integer, $m_0 = np$.

Indication: consider the ratio

$$\frac{P(\mu_n = m+1)}{P(\mu_n = m)} = \frac{(n-m)p}{(m+1)q}, \; [q = 1-p].$$

8) The encounter problem. […] [See here Sheynin, end of § 1.1.2]

Chapter 3. Numerical Characteristics of the Distributions of Probabilities

While dealing with discrete or continuous random variables, it is not always advisable to use tables or densities of distributions. The former can be insufficiently precise and the latter not precisely known whereas calculations are often complicated or cumbersome. However, many important problems can be solved by a few averaged characteristics of distributions, so we begin by averaging.

3.1. Averaging. Expectation of random variables

Consider the simplest notion of arithmetic mean. Suppose that we have a population of N elements with differing magnitudes of some indication x; for example, a batch of bulbs differing in the period of work or rainy days in a year differing by the amount of rainfall.

The arithmetic mean of indication x in a population is the ratio of the sum of the values of those indications in the population by the total number of its elements.

Denote by $x_1, x_2, \ldots, x_v$ the differing values of the studied indication, by M_k, the number of elements having indication x_k ($k = 1, 2, \ldots, v$) and let $N = M_1 + M_2 + \ldots + M_v$ be the total number of the elements. Then the arithmetic mean is

$$\overline{x} = \frac{x_1 M_1 + x_2 M_2 + \ldots + x_v M_v}{N} = x_1 \frac{M_1}{N} + x_2 \frac{M_2}{N} + \ldots + x_v \frac{M_v}{N}. \quad (3.1)$$

The arithmetic mean thus only depends on the relative magnitudes $M_1/N, M_2/N, \ldots, M_v/N$ rather than on $M_1, M_2, \ldots, M_v$.

We turn now to random variables and begin with a discrete variable of a special type. Choose randomly an element; this is best imagined, just like in § 5.1.2, by extracting balls from an urn. The magnitude y of the selected element is a discrete random variable ξ with

values: $x_1, x_2, \ldots, x_v$; probabilities: $M_1/N, M_2/N, \ldots, M_v/N$. (3.2)

The arithmetic mean of the indications is here as though the mean *expected* value of ξ. And so, the expectation of ξ is

$$\mathrm{E}\xi = x_1(M_1/N) + x_2(M_2/N) + \ldots + x_v(M_v/N).$$

Here, however, the sum ought to be interpreted as the sum of the products of the values of ξ by their probabilities which allows us to extend at once the notion of expectation on any discrete variable ξ with values $x_1, x_2, \ldots, x_v$ and probabilities $p_1, p_2, \ldots, p_v$.

Definition 1. Expectation $\mathrm{E}\xi$ of a discrete random variable ξ is the sum of the products of all of its possible values $\{x_k\}$ by their probabilities $\{p_k\}$:

$$\mathrm{E}\xi = x_1p_1 + x_2p_2 + \ldots = \sum x_kp_k. \quad (3.3)$$

If there are infinitely many such values, we will assume that the series (3.3) absolutely converges, otherwise Eξ does not exist; we will not consider such cases. Now we may extend the notion of expectation on continuous random variables by replacing p_k by the differential of probability $dP_x = \varphi(x)dx$.

Definition 2. Expectation Eξ of a continuous random variable ξ is the integral of the product of its values by the density of distribution $\varphi(x)$:

$$\mathrm{E}\xi = \int x\varphi(x)dx. \quad (3.4)$$

The integral is taken over all the interval of the possible values of ξ. It is often written as if that interval is $(-\infty, \infty)$ even if the possible values of ξ only cover a finite interval. In such cases we assume that beyond that interval $\varphi(x) = 0$. If, however, the domain of ξ covers the entire numerical axis, the improper integral is assumed to converge absolutely; again, the expectation does not otherwise exist and we will not consider such cases.

It is important to note that all the properties of expectation (or, more precisely, of the very operation of averaging) are quite identical for discrete and continuous random variables. It is also possible to provide a single definition of expectation for any random variable, but, just like in § 2.3.4, it will require the knowledge of the Stieltjes integral.

3.1.1. *The properties of the expectation.* The most important property of averaging is linearity: the expectation of a linear combination of random variables is [the same] linear combination of their expectations:

$$\mathrm{E}(C_1\xi_1 + C_2\xi_2 + \ldots + C_n\xi_n) = C_1\mathrm{E}\xi_1 + C_2\mathrm{E}\xi_2 + \ldots + C_n\mathrm{E}\xi_n \quad (3.5)$$

where $C_1, C_2, \ldots, C_n$ are constants.

For proving this property, suffice it to prove the following theorems.

1) A constant factor can be taken out of the sign of expectation

$$\mathrm{E}C\xi = C\mathrm{E}\xi. \quad (3.6)$$

2) The expectation of a sum of two random variables is equal to the sum of their expectations (the addition theorem for expectations):

$$\mathrm{E}(\xi + \eta) = \mathrm{E}\xi + \mathrm{E}\eta. \quad (3.7)$$

Formula (3.6) is especially easy to prove for a discrete random variable ξ whose multiplication by a constant C is determined by table (2.6):

$$\mathrm{E}C\xi = \sum Cx_kp_k = C\sum x_kp_k = C\mathrm{E}\xi.$$

See below the proof for continuous variables.

We will prove the addition theorem (3.7) for continuous random variables. Denote by φ(*x, y*) the density of the joint distribution [of ξ and η] and by χ(*z*) the density of their sum, ς = ξ + η. Then by formulas (3.4) and (2.39)

$$\mathrm{E}\varsigma = \int z\chi(z)dz = \int_{-\infty}^{\infty} zdz \int_{-\infty}^{\infty} \varphi(x, z-x)dx.$$

Change the order of integration and replace *z* by *x* + *y*:

$$\mathrm{E}(\xi + \eta) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} z\varphi(x, z-x)dz = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} (x+y)\varphi(x+y)dy. \quad (3.8)$$

Recall the linearity of the integral and apply formulas (2.35) and (2.36):

$$\mathrm{E}(\xi + \eta) = \int_{-\infty}^{\infty} xdx \int_{-\infty}^{\infty} \varphi(x, y)dx + \int_{-\infty}^{\infty} ydy \int_{-\infty}^{\infty} \varphi(x, y)dx =$$

$$\int_{-\infty}^{\infty} x\psi_1(x)dx + \int_{-\infty}^{\infty} y\psi_2(y)dy = \mathrm{E}\xi + \mathrm{E}\eta.$$

For discrete random variables the proof is similar. At first, it is easy to be convinced in that the expectation of (ξ + η) can be calculated by the formula

$$\mathrm{E}(\xi + \eta) = \sum(x_k + y_l)p_{kl} = \sum x_k p_{kl} + \sum y_l p_{kl},$$

where the sums cover all values of x_k and y_l. By formula (1.20), if $p_k = P(\xi = x_k)$, the first sum is

$$\sum x_k(p_{k1} + p_{k2} + \ldots) = \sum x_k p_{kl} = \mathrm{E}\xi.$$

In a similar way $\sum y_l p_{kl} = \mathrm{E}\eta$, QED.

3) The expectation of a constant (non-random) magnitude *C* is that very magnitude. Indeed, *C* can be considered a random variable with a single possible value *C* and probability 1, so that E*C* = *C*.

4) The expectation of a product of *independent* random variables is the product of their expectations (the multiplication theorem for expectations):

$$\mathrm{E}\xi\eta = \mathrm{E}\xi\mathrm{E}\eta. \quad (3.9)$$

Here is a proof for discrete variables. The distribution of the product ξη has table

$$x_1y_1, x_1y_2, x_2y_1, x_2y; \ldots; p_1q_1, p_1q_2, p_2q_1, p_2q_2, \ldots$$

Equal values should be combined as also the probabilities (cf. the determination of the sum of discrete variables in § 2.1.1). We can therefore write the expectation of ξη as

$$\text{E}\xi\eta = \sum x_k y_l p_k q_l$$

where the sum covers all possible values of x_k and y_l of ξ and η respectively. We can write this equality as

$$\text{E}\xi\eta = \sum x_k p_k \sum y_l q_l, \text{ QED.}$$

For continuous random variables the proof is easily carried out by formula (3.13) below but we leave it for the readers. Formula (3.9) is not difficult to extend on any number of mutually independent multipliers.

3.1.2. *Calculation of expectations of functions*

1) Let ξ be a discrete random variable taking values x_k with probabilities p_k. The function $f(\xi)$ is again a discrete variable and its expectation is

$$\text{E}f(\xi) = \sum f(x_k)P[f(\xi) = f(x_k)] \qquad (3.10)$$

where the sum covers all the different values of $f(x_k)$.

It occurs that the expectation of $f(\xi)$ can be calculated without deriving the distribution of its probabilities but directly by the distribution of ξ itself. Indeed,

$$\text{E}f(\xi) = \sum f(x_k)p_k \qquad (3.11)$$

where the sum covers all the values x_k of ξ.

Before proving (3.11) in the general case, we note that if all those values are different, function $f(x)$ has the table of distribution (2.27) so that formula (3.11) coincides with (3.10). In the general case equal numbers can occur in the values of $f(x_k)$. Suppose for the sake of definiteness that only two values are equal, $f(x_1) = f(x_2)$. Then the probability of the event $f(\xi) = f(x_1)$ is $p_1 + p_2$ and formula (3.10) can now be written as

$$f(x_1)P[f(\xi) = f(x_1)] = f(x_1)(p_1 + p_2) = f(x_1)p_1 + f(x_2)p_2$$

and (3.11) follows once more.

2) Expectation of a function $f(\xi)$ of a continuous random variable ξ can also be calculated directly by issuing from the density $\varphi(x)$ of the distribution of ξ itself

$$\text{E}f(\xi) = \int f(x)\varphi(x)dx. \qquad (3.12)$$

We will only prove this formula for an increasing $f(x)$. Denote the density of the distribution of $\eta = f(\xi)$ by $\psi(y)$ and replace y by $f(x)$ in the formula for expectation

$$E\eta = \int y\psi(y)dy.$$

Then, by formula (2.28) we have $\psi(y)dy = \varphi(x)dx$ and immediately arrive at (3.12). Thus, for function $C\xi$ (3.12) provides

$$EC\xi = \int C\varphi(x)dx = C\int x\varphi(x)dx = CE\xi$$

which proves formula (3.6) for continuous variables.

3) Here without proof are the respective formulas for calculating expectations of functions of two variables. For discrete variables

$$Ef(\xi, \eta) = \sum f(x_k, y_l)p_{kl}$$

where the summation is over all the values of x_k and y_l of ξ and η and p_{kl} is the probability of random events $\xi = x_k$ and $\eta = y_l$.

For continuous variables

$$Ef(\xi, \eta) = \iint f(x, y)\varphi(x, y)dxdy \qquad (3.13)$$

where $\varphi(x, y)$ is the density of distribution of the random point (ξ, η). Above, see for example formula (3.8), we have discussed particular cases of these formulas taking $f(x, y) = x + y$ and xy.

3.2. The centre of the distribution of a random variable

The expectation of a random variable provides a convenient characteristic of its whereabouts. It has the same dimensionality as its values and is located within their possible interval. Thus, if all the values of a random variable ξ are within interval (α_1, α_2),

$$P(\alpha_1 < \xi < \alpha_2) = \int_{\alpha_1}^{\alpha_2} \varphi(x)dx = 1$$

and inequalities

$$\int_{\alpha_1}^{\alpha_2} \alpha_1\varphi(x)dx < \int_{\alpha_1}^{\alpha_2} x\varphi(x)dx < \int_{\alpha_1}^{\alpha_2} \alpha_2\varphi(x)dx$$

[naturally] lead to

$$\alpha_1 < \xi < \alpha_2.$$

In particular, if all the values of ξ are positive, $E\xi$ is also positive. It will be shown that the arithmetic means of the sample values of a random variable group around its expectation, see next chapter. The following definition stresses the role of expectation. Unlike the operation of averaging in itself, it is the main characteristic of the location of a random variable. *The centre of the distribution of probabilities of a random variable is its expectation*[6].

Example 1. Suppose that ξ is the number of expended cartridges when firing as in Example 2 of § 1.2. By the table of distribution (2.3) we find that the expectation of that number is

$$\mathrm{E}\xi = 1\cdot 0.8 + 2\cdot 0.16 + 3\cdot 0.04 = 1.24.$$

It is not an integer. For showing the practical usefulness of that calculation let us imagine that that trial was made a hundred times and denote the number of expended cartridges at trial k by ξ_k. Then

$$\xi_1 + \xi_2 + \ldots + \xi_{100}$$

is the total number of the expended cartridges. Taking $\xi_k = 1.24$, $k = 1, 2, \ldots, 100$, we have, bearing in mind the linearity property,

$$\mathrm{E}\xi = \mathrm{E}\xi_1 + \mathrm{E}\xi_2 + \ldots + \mathrm{E}\xi_{100} = 124.$$

2) The centre for the Poisson distribution (2.5)

$$\mathrm{E}\xi = \sum_{m=0}^{\infty} m\frac{a^m}{m!}e^{-a} = ae^{-a}(1+a+\frac{a^2}{2!}+\ldots+\frac{a^{m-1}}{(m-1)!}+\ldots) = a.$$

This explains the meaning of the parameter a: it is the expectation of random variable ξ having the Poisson distribution (2.5).

3) The centre of distribution of frequency μ_n and relative frequency w_n of a random event. A direct calculation of the expectation by the table of distribution (2.13) leads to

$$\mathrm{E}\mu_n = \sum_{m=0}^{n} mC_n^m p^m q^{n-m}.$$

For speeding up calculations we make use of the linearity of the expectation and formulas (2.9) and (2.10) providing expressions of the expectations of random variables μ_n and w_n by indicator variables $\lambda_1, \lambda_2, \ldots, \lambda_n$. A direct calculation by the tables of distribution (2.8) gives

$$\mathrm{E}\lambda_k = 1\cdot p + 0\cdot q = p,\ k = 1, 2, \ldots, n. \qquad (3.14)$$

Therefore, *the centre of distribution of an indicator variable is the probability of the studied event*. Then,

$$\mathrm{E}\mu_n = \mathrm{E}\lambda_1 + \mathrm{E}\lambda_2 + \ldots + \mathrm{E}\lambda_n = np, \qquad (3.15)$$
$$\mathrm{E}w_n = (1/n)\mathrm{E}\mu_n = p. \qquad (3.16)$$

The centre of distribution of the relative frequency w_n of a random event is its probability in a single trial and the centre for the frequency μ_n is n times larger.

These conclusions agree with our intuitive idea about expectation. If, for example, the probability of a random event is $p = 0.2$, and 100 trials are made, we expect $np = 20$ occurrences of that event. Or, if the probability of a substandard manufactured article in a large batch is $p = 1\%$, then in a sample of $n = 1000$ articles we tend to expect $np = 10$ such articles. We certainly admit the possibility of some deviations, but here we are discussing the mean expected results.

Note also that the linearity of the expectation allows us to derive a more general result than (3.16) from (3.14). If random event A has probability p_k in trial k, the centre of the distribution of the relative frequency w_n of A will be

$$\mathrm{E}w_n = (1/n)(\mathrm{E}\lambda_1 + \mathrm{E}\lambda_2 + \ldots + \mathrm{E}\lambda_n) = (1/n)(p_1 + p_2 + \ldots + p_n)$$

where n is the number of the trials. It will then be equal to the arithmetic mean of all the probabilities of A.

4) If random variable ξ is uniformly distributed on interval (α_1, α_2), its centre of distribution coincides with the midpoint of that interval. Indeed, the density of the uniform distribution is constant in that interval and equals $1/(\alpha_1 - \alpha_2)$ so that

$$\mathrm{E}\xi = \int_{\alpha_2}^{\alpha_1} \frac{xdx}{\alpha_2 - \alpha_1} = \frac{1}{\alpha_2 - \alpha_1}\frac{\alpha_2^2 - \alpha_1^2}{2} = \frac{\alpha_2 + \alpha_1}{2}.$$

5) *The centre of the normal distribution.* The centre of the simplest normal distribution (2.21) is zero since the density $\varphi_0(x)$ is an even function. And a random variable ξ obeying the general normal distribution can be expressed by a random variable ξ_0 (§ 2.4.1):

$$\xi = a + \sigma\xi_0 \text{ and } \mathrm{E}\xi = a + \sigma\mathrm{E}\xi_0 = a.$$

The centre a of the general normal distribution is its parameter. This statement ascertains the meaning of that parameter and indeed agrees with the symmetry of the normal curve of distribution with respect to straight line $x = a$.

Remark. If that curve is symmetric with respect to some line $x = a$, the centre of distribution invariably coincides with point a.

3.3. Characteristics of the scattering of a random variable. Notion of the moments of distribution

The scattering of random variable ξ is connected with the deviation $\xi - a$ from its centre of distribution $a = \mathrm{E}\xi$. A direct averaging of that deviation will not provide any numerical characterisation of the scattering since

$$\mathrm{E}(\xi - a) = \mathrm{E}\xi - a = 0.$$

In the mean, deviations of contrary signs mutually compensate each other. The main numerical characteristic of the scattering of a random variable ξ is the mean square deviation σ:

$$\sigma = \sigma(\xi) = \sqrt{\mathrm{E}(\xi - a)^2},\ a = \mathrm{E}\xi. \qquad (3.17)$$

The magnitude $E(\xi - a)^2 = \sigma^2(\xi) = \text{var}\xi$, is the *variance of* ξ. In accordance with formulas (3.11) and (3.12), the variances of discrete and continuous random variables are

$$\sigma^2(\xi) = \sum(x_k - a)^2 p_k,\ \sigma^2(\xi) = \int(x - a)^2\varphi(x)dx.$$

It is seen that the mean square deviation has the same dimensionality as the values of the random variable. The special role of that deviation is discussed in detail below, mostly in Chapters 4 and 5. In particular, it will be shown that deviations of random variables many times exceeding σ from their centre of distribution do not practically occur. Here, however, we restrict our discussion to considering examples of the simplest properties of the mean square deviation.

3.3.1. *Main rules of computing mean square deviations and variances*

1) If ξ is a random variable and *C* is constant,

$$\sigma(C\xi) = |C|\sigma(\xi),\ \sigma(\xi + C) = \sigma(\xi). \qquad (3.18), (3.19)$$

The following formulas are proved by direct calculations of variance:

$$\sigma^2(C\xi) = E(C\xi - EC\xi)^2 = E(C\xi - Ca)^2 = C^2E(\xi - a)^2 = C^2\sigma^2(\xi).$$
$$\sigma^2(\xi + C) = E[(\xi + C) - E(\xi + C)]^2 =$$
$$E(\xi + C) - (a + c)]^2 = E(\xi - a)^2 = \sigma^2(\xi).$$

2) For *independent* random variables the variance of their sum is the sum of their variances:

$$\sigma^2(\xi + \eta) = \sigma^2(\xi) + \sigma^2(\eta) \qquad (3.20)$$

(addition theorem for variances) and therefore

$$\sigma(\xi + \eta) = \sqrt{\sigma^2(\xi) + \sigma^2(\eta)}.$$

Proof of formula (3.20). Denote $E\xi = a$, $E\eta = b$, then $E(\xi + \eta) = a + b$ and

$$\sigma^2(\xi + \eta) = E[(\xi + \eta) - (a + b)]^2 =$$
$$E[(\xi - a)^2 + 2(\xi - a)(\eta - b) + (\eta - b)^2] =$$
$$E(\xi - a)^2 + 2E(\xi - a)(\eta - b) + E(\eta - b)^2.$$

For independent ξ and η

$$E(\xi - a)(\eta - b) = E(\xi - a)E(\eta - b) = 0$$

since, see above, $E(\xi - a) = 0$ [$E(\eta - b) = 0$ as well], QED.

The addition theorem for variances is easily generalized on any number of pairwise independent random variables.

Corollary. The variance of a linear combination of pairwise independent random variables $\xi_1, \xi_2, \ldots, \xi_n$ can be calculated according to formula

$$\sigma^2(C_1\xi_1 + C_2\xi_2 + C_n\xi_n) = C_1^2\sigma^2(\xi_1) + C_2^2\sigma^2(\xi_2) + \ldots + C_n^2\sigma^2(\xi_n)$$

which directly follows from formulas (3.20) and (3.18). And, if ξ_1, $\xi_2, \ldots, \xi_n$ have identical variances,

$$\sigma^2(\xi_k) = \sigma^2,\ k = 1, 2, \ldots, n$$

the variance of their arithmetic mean is

$$\sigma^2[\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n}] = \frac{\sigma^2(\xi_1) + \sigma^2(\xi_2) + \ldots + \sigma^2(\xi_n)}{n^2} = \frac{\sigma^2}{n}.$$

The mean square deviation of that mean is therefore

$$\sigma[\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n}] = \frac{\sigma}{\sqrt{n}}, \qquad (3.21)$$

a very important formula for treating observations (see Chapter 6).

3.3.2. *Examples*

1) *The mean square deviation of the relative frequency*. Formula (2.10) shows that the relative frequency w_n is an arithmetic mean of mutually independent indicator random variables $\lambda_1, \lambda_2, \ldots, \lambda_n$ having identical tables of distribution (2.8) $\lambda_k = 1, 0$ with probabilities p and q, $p + q = 1$, $k = 1, 2, \ldots, n$:

$$w_n = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_n}{n}.$$

Now directly calculate the variance of λ_k bearing in mind that the centre of distribution of its probabilities is p:

$$\sigma^2(\lambda_k) = E(\lambda_k - p)^2 = (1 - p^2)p + (0 - p^2)q = q^2p + p^2q = pq.$$

The mean square deviation is therefore

$$\sigma(\lambda_k) = \sqrt{pq},\ k = 1, 2, \ldots, n.$$

From (3.21) we have now

$$\sigma(w_n) = \frac{\sqrt{pq}}{\sqrt{n}}. \qquad (3.22)$$

and by formula (3.18)

$\sigma(\mu_n) = \sigma(nw_n) = n\sigma(\mu_n) = \sqrt{npq}$.

2) *The mean square deviation of a random variable* ξ *uniformly distributed on interval* (α_1, α_2). We have calculated the centre of the distribution

$a = \mathrm{E}\xi = (\alpha_1 + \alpha_2)/2$.

Directly calculate now the variance:

$$\sigma^2(\xi) = \mathrm{E}[\xi - \frac{\alpha_1 + \alpha_2}{2}]^2 = \int_{\alpha_1}^{\alpha_2} [x - \frac{\alpha_1 + \alpha_2}{2}]^2 \frac{dx}{\alpha_1 - \alpha_2} = \frac{(\alpha_2\text{-}\alpha_1)^2}{12},$$

$$\sigma(\xi) = \frac{\alpha_2 - \alpha_1}{2\sqrt{3}}.$$

This mean square deviation is proportional to the length of the interval (α_1, α_2) and is approximately equal to 1/3 of it.

3) *The variance of the normal distribution*. Let random variable ξ_0 have the simplest normal distribution (2.21) with centre $\mathrm{E}\xi_0 = 0$. The variance of the distribution is therefore

$$\mathrm{var}\xi_0 = \mathrm{E}\,\xi_0^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp[-\frac{x^2}{2}] dx = 1. \qquad (3.23)$$

The integral can be conveniently calculated by parts. For random variable $\xi = a + \sigma\xi_0$ having the general normal distribution (2.25) the variance is

$\mathrm{var}\xi = \mathrm{var}(a + \sigma\xi_0) = \sigma^2\mathrm{var}\xi_0 = \sigma^2$.

It follows that $\sigma(\xi) = \sigma$. Together with the formula $\mathrm{E}\xi = a$ (see above) it completely ascertains the meaning of the parameters a and σ of the general normal distribution (2.25): a is the centre of the distribution and σ^2, the variance.

3.3.3. *The minimality property of the centre. The mean square deviation of random variable* ξ *from the centre of distribution* $a = \mathrm{E}\xi$ *is smaller than the same variation from any other number*:

$\mathrm{E}(\xi - a)^2 < \mathrm{E}(\xi - C)^2,\ C \neq a.$

Proof. Since $\mathrm{E}(\xi - a) = 0$,

$$\mathrm{E}(\xi - C)^2 = \mathrm{E}[(\xi - a) + (a - C)]^2 =$$
$$\mathrm{E}(\xi - a)^2 + 2(a - C)\mathrm{E}(\xi - a) + (a - C)^2 = \mathrm{E}(\xi - a)^2 + (a - C)^2 \quad (3.24)$$

and

$\mathrm{E}(\xi - a)^2 \leq \mathrm{E}(\xi - C)^2$

with the equality only taking place when $C = a$.

Formula (3.24) is often applied for calculating variances; note its similarity with the corresponding theorem about the moments of inertia. When $C = 0$ this formula provides

$$\sigma^2(\xi) = \mathrm{E}\xi^2 - a^2. \qquad (3.25)$$

So let us calculate the variance of the Poisson distribution (2.5). It is easiest to begin with $\mathrm{E}\xi^2$:

$$\mathrm{E}\xi^2 = \mathrm{E}[\xi(\xi - 1)] + \mathrm{E}\xi =$$

$$\sum_{m=0}^{\infty} m(m-1)\frac{a^m}{m!}e^{-a} + a = a^2 e^{-a}\sum_{m=2}^{\infty}\frac{a^{m-2}}{(m-2)!} + a = a^2 + a.$$

Now the variance:

$$\sigma^2(\xi) = \mathrm{E}\xi^2 - a^2 = (a^2 + a) - a^2 = a.$$

It is useful to note that for the Poisson distribution both its centre and variance coincide with the value of its parameter.

3.3.4. *Notion of the moments of distribution*. The two main characteristics of distributions, their centres $\mathrm{E}\xi = a$ and variances $\mathrm{E}(\xi - a)^2 = \sigma^2$, are particular cases of the moments of distribution which Chebyshev introduced for investigating the laws of distribution.

The k-th *initial* moment is the expectation of the k-th power of a random variable, $\mathrm{E}\xi^k$. The k-th *central* moment is the expectation of the k-th power of the deviation of a random variable from the centre of its distribution, $\mathrm{E}(\xi - a)^k$. There exist simple connections between these moments, and they are easily established by the Newtonian binomial. For example,

$$\mathrm{E}(\xi - a)^2 = \mathrm{E}\xi^2 - 2a\mathrm{E}\xi + a^2 = \mathrm{E}\xi^2 - a^2.$$
$$\mathrm{E}(\xi - a)^3 = \mathrm{E}\xi^3 - 3a\mathrm{E}\xi^2 + 3a^2\mathrm{E}\xi - a^3 = \mathrm{E}\xi^3 - 3a\mathrm{E}\xi^2 + 2a^2.$$

The former formula coincides with (3.25). As stated above, the first and the second moments, $\mathrm{E}\xi$ and $\mathrm{E}(\xi - a)^2$, characterize the centre of the location and the scattering of the random variable ξ. The third central moment, $\mathrm{E}(\xi - a)^3$, is applied for characterizing *the asymmetry of the distributions*. If the curve of distribution is asymmetric with respect to the straight line $x = a$, the third central moment (and all odd central moments) will disappear. Indeed, the density of distribution $\psi(y)$ of the random variable (of the deviation) $\eta = (\xi - a)$ will be an even function and all products $y^{2k+1}\psi(y)$ will be odd functions.

Therefore, if the third moment is not zero, the distribution cannot be symmetric. The magnitude of asymmetry is usually defined by its dimensionless coefficient

$$C_1 = \mathrm{E}(\xi - a)^3/\sigma^2(\xi).$$

Its sign indicates the direction of asymmetry. Moments higher than the third do not occur in elementary problems or the simplest applications of probability theory.

3.4. Exercises

1) Calculate the expectation of the product of indicator variables λ_1 and λ_2 as introduced in the example of § 2.2. Check that in this case the multiplication theorem for expectations is not applicable. *Answer*:

$$\mathrm{E}(\lambda_1\lambda_2) = 25/100\cdot24/99;\ \mathrm{E}\lambda_1\mathrm{E}\lambda_2 = (25/100)^2.$$

2) Prove by formula (3.13) the multiplication theorem for expectations of independent continuous random variables.

Indication. Apply the condition of independence $\varphi(x, y) = \psi_1(x)\ \psi_2(y)$ and express the double integral

$$\iint xy\psi_1(x)\psi_2(y)dxdy$$

as two ordinary integrals.

3) Find the centre and the mean square deviation for the distribution (2.2) of the number of points on a die. *Answer*:

$$\mathrm{E}\xi = \frac{1+2+3+4+5+6}{6} = 3.5;\ \sigma = \sqrt{\frac{35}{12}} = 1.71.$$

4) Solve the same problem for the points on two dice. *Answer*:

$$\mathrm{E}\xi = 7;\ \ \sigma = \sqrt{\frac{70}{12}} = 2.42.$$

Indication: Apply the addition theorems for expectations and variances.

5) Determine the expectation of the number of white balls when trials are made as in Exercise 2 in § 2.5. *Answer*: $\mathrm{E}\xi = 5/6$.

Indication: Express ξ as a sum of indicator variables connected with the extraction of each ball.

6) Determine the centre and the variance of distribution (2.4) of the number of expended cartridges as in Example 3 of § 2.1. Consider a numerical example for $p = 1/10$ and interpret the expectation. *Answer*:

$$\mathrm{E}\xi = \sum_{n=1}^{\infty} n(1-p)^{n-1}p = 1/p;$$

$$\sigma^2 = \mathrm{E}\xi^2 - (\mathrm{E}\xi)^2 = \sum_{n=1}^{\infty} n^2(1-p)^{n-1}p - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

If $p = 1/10$, $\mathrm{E}\xi = 10$. If the probability of a hit is each time 1/10, the first hit will be achieved in the mean after 10 attempts.

Indication: Apply the power series for $(1 - q)^{-2}$ and $(1 - q)^{-3}$.

7) Determine the centre and variance for the Pearson distribution (2.26). *Answer*:

$$\text{E}\xi = \int_0^\infty x \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{\beta^\alpha \Gamma(\alpha+1)}{\Gamma(\alpha)\beta^{\alpha+1}} = \frac{\alpha}{\beta};$$

$$\sigma^2 = \text{E}\xi^2 - (\text{E}\xi)^2 = \frac{\beta^\alpha \Gamma(\alpha+2)}{\Gamma(\alpha)\beta^{\alpha+2}} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}.$$

Indication: Apply integration by parts or the main property of the gamma-function.

8) Prove the addition theorem for the third central moments of independent random variables ξ and η:

$$\text{E}[(\xi + \eta) - (a + b)]^3 = \text{E}(\xi - a)^3 + \text{E}(\eta - b)^3.$$

9) Find the coefficient of asymmetry for the binomial distribution of the frequency μ_n. *Answer*:

$$\frac{\text{E}(\mu_n - np)^3}{\sigma^3(\mu_n)} = \frac{npq(q-p)}{(npq)^{3/2}} = \frac{q-p}{\sqrt{npq}}.$$

Indication: At first calculate the third central moment for the indicator variable λ:

$$\text{E}(\lambda - p)^3 = (1 - p)^3 p + (0 - p)^3 q = pq(q - p).$$

Then apply the addition theorem for the third central moments.

10) Prove that the coefficient of asymmetry for the Pearson distribution (2.26) is twice larger than the so-called coefficient of variation

$$C_v = \sigma_\xi/\text{E}\xi = 1/\surd\alpha.$$

Indication. When calculating the central moment $\text{E}(\xi - a)^3$ express it through the initial moments.

11) Calculate the fourth moment for the general normal distribution (2.25). *Answer*:

$$\text{E}(\xi - a)^4 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-a)^4 \exp[-\frac{(x-a)^2}{2\sigma^2}]dx = 3\sigma^4.$$

Indication: replace $(x - a)$ by $t\sigma$ and integrate by parts.

Chapter 4. The Law of Large Numbers

It is impossible to foresee the value of a random variable in a trial. However, the behaviour of the sum of a large number of random variables almost looses randomness and becomes regular. Necessity carves its way through a multitude of chances and the pertinent theorems are known by the generic name law of large numbers.

4.1. Random events with very low probabilities

To recall, the probability p of a random event is a number objectively characterizing, under specified conditions, the possibility of its occurrence. The event's relative frequency is a random variable with a distribution of probabilities having at its centre that same p (§ 3.2, Example 3). The value of the probability cannot be directly derived from an experiment but each repetition of n trials provides a definite experimental value of relative frequency. In the beginning of this book I have indicated that, given a sufficiently large number of trials n, its value is, as a rule, very near to probability p.

This general conclusion connects theory and practice, but it is too indefinite for numerical estimates: we know that the relative frequency is near to probability, but are unable to say just how near it is. Therefore, we will now issue from a narrower but more definite principle concerning events having very low probabilities. Such events occur extremely seldom. If, for example, an event has probability 0.000001, it happens approximately once in a million trials. However, this certainly does not mean that it occurs in the millionth trial; it can happen in one of the first of them.

Experience convinces us in that as a rule, given a small number of trials, such rare events do not happen at all. Thus, having a ticket of a lottery in which only 1 prize is won for every million tickets, you will hardly hope to be lucky (although someone will actually win!). But if there are only 500,000 or 10,000 tickets? A question appears: how low should the probability of a random variable be for neglecting its appearance in a single trial? The theory of probability cannot say anything here since this question belongs to its practical applications. Here are examples of two events.

1) Automatically manufacturing articles; the probability of obtaining an article of a non-standard size is 0.01 and the sizes will be checked. If these articles are not expensive, it is quite possible to abstain from checking all of them, i. e., to neglect the probability of 0.01[7].

2) The same problem concerning parachutes. It is certainly inadmissible to neglect probability 0.01 that the parachute will not open. Each should be checked.

A certain boundary of very low probability is assigned in each field of the application of probability theory. This boundary is established according to the principle of practical impossibility of unlikely events. It is assumed that an event having a probability lower than that boundary will not occur in a single trial.

This principle will ascertain the practical meaning of the theorem discussed below. It is sometimes called the principle of practical certainty [of the contrary event].

An important remark is in order. Suppose that, when issuing from some hypothesis, we find that the probability of event A is lower than the assigned boundary, but that it nevertheless occurred in a single trial. It will then be reasonable to question our hypothesis and look for a non-random cause of A. This is especially clearly expressed in a venerable story (Bertrand 1888, pp. VII – VIII). A man undertook to cast three sixes with three dice, and indeed achieved it. You will say that such an outcome was possible, but he continued to be successful 2, 3, 4 and 5 times in succession. *What the hell*, cried his adversary, *the dice are loaded*! And so they were.

The probability of casting three sixes is $1/6^3 = 1/216$; for 5 successes it should be raised to the power of 5.

4.2. The Jakob Bernoulli theorem and the stability of relative frequencies

Suppose that random variable *A* has probability *p* of appearing in a trial and that *n* such trials are made. We know that the relative frequency of *A* will be the random variable w_n whose centre of distribution coincides with *p*. Its mean square deviation will decrease with the increase of *n*, see formula (3.22):

$$\sigma(w_n) = \frac{\sqrt{pq}}{\sqrt{n}}.$$

It follows that, as the number of trials increases, the values of the relative frequency of a random event will scatter ever less, they will ever nearer group around the probability of that event. The remarkable Jakob Bernoulli theorem published in 1713 specifies this proposition.

If the probability p of a random event remains invariable in a sequence of n independent trials, the probability that the deviation of the relative frequency w_n of the event from p exceeds a given number $\varepsilon > 0$ *tends to disappear with an unbounded increase of n*:

$$\lim P(|w_n - p| > \varepsilon) = 0,\ n \to \infty. \qquad (4.1)$$

With a sufficiently large *n* that probability will thus become lower than the assigned boundary of very low probabilities (see § 4.1), and it is practically certain that the inequality in (4.1) will not happen, and that, consequently, the contrary inequality will be obeyed:

$$|w_n - p| \le \varepsilon. \qquad (4.2)$$

The Bernoulli theorem can also be therefore formulated thus:

A sufficiently large number of trials ensures a practical certainty that the deviation of the relative frequency of a random event from its probability will not exceed in absolute value any however small and given beforehand ε.

This theorem is a very particular case of the Chebyshev theorem (§ 4.3)[8]. Note that however large is *n*, we cannot categorically maintain that the inequality (4.2) will invariably take place; we are only practically certain in its fulfilment. For stressing the distinction of this proposition from the usual notion of limit we sometimes introduce a special notion of *limit in probability*.

If random event *A* occurred *m* times in *n* trials, *m/n* is a particular experimental value of w_n. With a sufficiently large *n* we may be practically certain in that the approximate equality

$$m/n \approx p \qquad (4.3)$$

will be satisfied as precisely as desired. In practice, this fact is manifested in that the values of *m/n* are stable; see the discussion of this term in the beginning of this book.

Equality (4.3) can serve for approximately calculating an unknown probability of a random event given statistical data. Thus, in the 19th century [and even earlier] it was established that the relative frequency of a male birth, 0.512, is stable. We can therefore conclude that the probability of that event has a probability near to 0.512. When given a certain number *n*, it is necessary to estimate the precision of the equality (4.3), see the next chapter.

4.3. The Chebyshev theorem

He proved it in 1867 for independent random variables. Consider a sequence of pairwise independent random variables $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ with any distributions of probabilities and suppose that they have definite expectations and variances

$$\mathrm{E}\,\xi_k = a_k,\ \mathrm{E}(\xi_k - a_k)^2 = \sigma_k^2,\ k = 1, 2, \ldots \qquad (4.4)$$

Calculate the arithmetic mean of the first *n* random variables:

$$\overline{\xi}_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n}. \qquad (4.5)$$

Its expectation is

$$\mathrm{E}\,\overline{\xi}_n = \frac{a_1 + a_2 + \ldots + a_n}{n} = \overline{a}_n. \qquad (4.6)$$

The variance of $\overline{\xi}_n$ is not equal to the arithmetic mean of the variances but *n* times less:

$$\sigma^2(\overline{\xi}_n) = \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n^2}. \qquad (4.7)$$

Suppose that $\sigma_k^2 \le H$, $k = 1, 2, \ldots$ Then, as $n \to \infty$, the variance of the arithmetic mean tends to zero since $\sigma^2(\overline{\xi}_n) \le H/n$. It follows that, as *n* increases, the values of $\overline{\xi}_n$ will scatter ever less, will ever nearer group around the centre of its distribution. This can be interpreted as saying that the random deviations of both signs are partially compensated in the arithmetic mean.

Now let us estimate those possible deviations by the universal [Bienaymé –] Chebyshev inequality.

4.3.1. *The* [*Bienaymé –*] *Chebyshev inequality*. It estimates the probability that the deviation of any random variable ξ from the centre of its distribution $a = \mathrm{E}\xi$ exceeds a given positive number ε:

$$P(|\xi - a| > \varepsilon) < \frac{\sigma^2(\xi)}{\varepsilon^2}. \qquad (4.8)$$

This probability is the lower the less is the variance $\sigma^2(\xi)$.

We will prove this inequality for continuous random variables. By the main formula (2.19) we have

$$P(|\xi - a| > \varepsilon) = \int \varphi(x)dx.$$

The integral is taken over $|x - a| > \varepsilon$ or over $(-\infty, a - \varepsilon)$ and $(a + \varepsilon, \infty)$. In both these intervals

$$1 < \frac{(x-a)^2}{\varepsilon^2} \text{ and } \varphi(x) \leq \frac{(x-a)^2\varphi(x)}{\varepsilon^2}$$

so that the integral above is

$$\int \varphi(x)dx \leq \frac{1}{\varepsilon^2}\int (x-a)^2\varphi(x)dx.$$

It only remains to note that

$$\int (x-a)^2\varphi(x)dx \leq \int_{-\infty}^{\infty} (x-a)^2\varphi(x)dx = \sigma^2(\xi).$$

Here, in the left side, the integral is taken over $|x - a| > \varepsilon$. We recommend the reader to prove in a similar way the [Bienaymé –] Chebyshev inequality for discrete variables.

4.3.2. *The Chebyshev theorem.* Apply the inequality (4.8) to $\overline{\xi}_n$:

$$P(|\overline{\xi}_n - \overline{a}_n| > \varepsilon) < \frac{\sigma^2(\overline{\xi}_n)}{\varepsilon^2} < \frac{H}{n\varepsilon^2}. \qquad (4.9)$$

Be ε as small as desired, it is always possible to assign such a large n that the right side of this inequality also becomes as small as desired, lower than the boundary chosen for *very low* probabilities. We will then be practically certain that the inequality in the left side is not obeyed, that the contrary inequality is taking place:

$$|\overline{\xi}_n - \overline{a}_n| \leq \varepsilon.$$

Given a sufficiently large number of independent random variables, it will be practically certain that the deviation of their arithmetic mean from the centre of its distribution will not exceed an arbitrarily small and assigned beforehand number ε.

The law of large numbers consists in the slight scatter of $\overline{\xi}_n$ around the centre of its distribution if only n is a large number. The Chebyshev theorem provides the exact mathematical expression of this proposition:

The arithmetic mean $\overline{\xi}_n$ of the first terms of sequence $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ of pairwise independent random variables with restricted variances obeys the equality

$$\lim P(|\overline{\xi}_n - \mathrm{E}\overline{\xi}_n| > \varepsilon) = 0 \text{ as } n \to \infty. \qquad (4.10)$$

This formula follows from (4.9). Indeed, the right side of the latter tends to zero as $n \to \infty$ and the left side, being a probability, is not negative.

4.3.3. *A particular case of the Chebyshev theorem.* Suppose that all the random variables $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ have the same centre of distribution $\mathrm{E}\xi_k = a$, $k = 1, 2, \ldots, n, \ldots$ Then the centre of distribution of $\overline{\xi}_n$ will also be a:

$$\mathrm{E}\overline{\xi}_n = (1/n)[\mathrm{E}\xi_1 + \mathrm{E}\xi_2 + \ldots + \mathrm{E}\xi_n] = a.$$

Formula (4.10) becomes

$$\lim P(|\overline{\xi}_n - a| > \varepsilon) = 0. \qquad (4.11)$$

Khinchin (1927) proved that formula (4.11) takes place for independent identically distributed random variables without imposing any restrictions on σ (which can be infinite) if only a is finite.

4.3.4. *The proof of the Bernoulli theorem.* Choose the indicator variables as the ξ_k, then their arithmetic mean will be equal to the relative frequency of the random event

$$\overline{\lambda}_n = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_n}{n} = w_n.$$

Since $\mathrm{E}\lambda_k = p$, $\sigma^2(\lambda_k) = pq < 1$, formula (4.11) is transformed into (4.1); QED.

4.4. Stability of sample means and the method of moments

Consider at first the statistical problem of mean values. Suppose that n elements differing in some quantitative indication x are randomly chosen from a population of N. May we say that the arithmetic means of that indication in the sample and population are near? The Chebyshev theorem answers this question if only the sample is taken with replacement.

Connect each selected k-th element with random variable ξ_k, a possible value of the k-th indication. Since the sample is selected with replacement the choice is always made from the initial population and random variables $\xi_1, \xi_2, \ldots, \xi_n$ will be independent and have an identical distribution of the type of (3.2). As stated in § 3.2, the centre of the distribution of all these variables coincides with the arithmetic mean of the indication in the general population, i. e., with the so-called *general mean a*:

$$\mathrm{E}\xi_k = a, \; k = 1, 2, \ldots, n.$$

The Chebyshev theorem (4.11) is therefore valid for $\overline{\xi}$ and the mean possible sample value of the indication tends in probability to the general mean as the size of the sample unboundedly increases.

Practical conclusion: the experimental value of each random variable ξ_k is the value which we find in the k-th element of the sample; the experimental value of random variable $\overline{\xi}$ is the sample mean $\overline{x}$. For a sufficiently large n of a random sample made with replacement we can be practically certain in that the sample mean will arbitrarily little differ from the general mean and

$$\overline{x} \approx a. \qquad (4.12)$$

The sample means are therefore stable: for two random samples of a sufficiently large size taken with replacement they should approximately coincide. This conclusion well enough agrees with experience. How near is the sample mean to the general mean only depends on the sample size but not on its ratio to the size of the general population. Thus, for the same values of σ, a sample containing 1% of a million elements provides more precise information about the general mean than a 2%-sample from a thousand elements[9].

If the size of the general population is very large as compared with the sample size, replacement becomes insignificant, and the conclusion above can also be applied to samples taken without replacement. This is especially important for applications since the general mean is often unknown and has to be judged by the sample mean. For example, the mean life of a bulb from a large batch can only be ascertained by a random sample. The necessary estimate of precision in such cases is provided in Chapter 6.

4.4.1. *On the method of moments*. The approximate equality (4.12) can also be interpreted otherwise. Suppose that ξ is a random variable with a finite centre of distribution $E\xi = a$. In independent trials ξ takes the values $x_1, x_2, \ldots, x_n$ which can be considered as the values of *differing* random variables $\xi_1, \xi_2, \ldots, \xi_n$ with the same distribution of probabilities as ξ itself. They can be assumed independent since the trials were independent. Then $\overline{x}$ can be thought of as an experimental value of random variable $\overline{\xi}$ for which the Chebyshev theorem in the form (4.11) is valid. With a sufficiently large n we can therefore expect that the approximate equality

$$\overline{x} \approx E\xi = a \qquad (4.13)$$

is satisfied precisely enough. Therefore, *the approximate value of the expectation of a random variable is the arithmetic mean of its experimental values*.

This proposition allows us to determine approximately not only the centre but other moments of the distribution as well. For example, we arrive at an approximate formula

$$\sigma^2(\xi) = \mathrm{E}(\xi - a)^2 \approx \frac{\sum (x_k - a)^2}{n} \qquad (4.14)$$

where the sum covers all the experimental data $x_1, x_2, \ldots, x_n$. Indeed, the right side of the approximate equality can be considered as a particular value of the arithmetic mean of *n* independent and identically distributed random variables $(\xi_k - a)^2$ with expectation

$$\mathrm{E}(\xi_k - a)^2 = \mathrm{E}(\xi - a)^2 = \sigma^2(\xi),\ k = 1, 2, \ldots, n.$$

Therefore,

$$P[|\frac{(\sum \xi_k - a)^2}{n} - \sigma^2(\xi)| > \varepsilon] \to 0 \text{ as } n \to \infty.$$

Formula (4.14) includes the value of the usually unknown centre of distribution. It is therefore natural to replace it by its approximate value $\bar{x}$. However, unlike the approximate formulas (4.13) and (4.14), the thus derived formula

$$\sigma^2(\xi) \approx \frac{\sum (x_k - \bar{x})^2}{n} \qquad (4.15)$$

will not hold anymore. Although its right side can be considered as a particular value of the arithmetic mean of *n* random variables $(\xi_n - \bar{\xi}_n)^2$, their expectation will not now be $\sigma^2(\xi)$ because of the linear dependence between $\xi_1, \xi_2, \ldots, \xi_n$ and their mean $\bar{\xi}$. A direct calculation provides

$$\mathrm{E}(\xi_1 - \bar{\xi}_n)^2 = \mathrm{E}[(\xi_1 - a) - (\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} - a)]^2 =$$

$$\mathrm{E}[(\xi_1 - a)(1 - 1/n) - \frac{\xi_2 - a}{n} - \ldots - \frac{\xi_n - a}{n}]^2 =$$

$$\sigma^2(\xi_1)(1 - 1/n)^2 + \frac{\sigma^2(\xi_2) + \ldots + \sigma^2(\xi_n)}{n^2} = \frac{n-1}{n}\sigma^2(\xi)$$

and similarly for any *k*:

$$\mathrm{E}(\xi_k - \bar{\xi}_n)^2 = \frac{n-1}{n}\sigma^2(\xi).$$

The expectation is linear and the expectation of random variables

$$\frac{n}{n-1}(\xi_k - \bar{\xi}_n)^2$$

is this very $\sigma^2(\xi)$. Although these variables are not independent, the law of large numbers is valid for them just as well (we omit the proof). If only n is sufficiently large, the value of their arithmetic mean

$$\frac{1}{n}\sum\frac{n}{n-1}(\xi_k-\overline{\xi}_n)^2=\frac{1}{n-1}\sum(\xi_k-\overline{\xi}_n)^2$$

will little deviate from their centre $\sigma^2(\xi)$.

Formula (4.15) can therefore be corrected by introducing factor $n/(n-1)$ in its right side:

$$\sigma^2(\xi)\approx\frac{\sum(x_k-\overline{x})^2}{n-1}. \qquad (4.16)$$

The right side is the sample variance denoted b s_n^2. It is useful to note that for large values of n the correction is relatively small and formulas (4.15) and (4.16) do not practically differ. For small values of n the difference is however very noticeable. If only n is known, the errors of all the approximate formulas provided above should be estimated. Some estimates are given in Chapter 6.

The approximate derivation of the moments of the distribution by experimental data ensures the possibility of calculating the parameters of the distribution provided that its type is known. Here are examples.

1) The parameters a and σ^2 of the general normal distribution are its centre and variance (§ 3.3). They can be derived by formula (4.14) and (4.17) if only experimental data are given.

2) The unknown parameters of the uniform distribution can be the ends of the interval of possible values α_1 and α_2. The moments are (§ 3.3)

$$\alpha_1=\mathrm{E}\xi=\frac{\alpha_1+\alpha_2}{2},\ \sigma=\sigma(\xi)=\frac{\alpha_2-\alpha_1}{2\sqrt{3}},$$

therefore $\alpha_1=a-\sigma\sqrt{3}$, $\alpha_2=a+\sigma\sqrt{3}$ and formulas (4.14) and (4.17) provide a and σ^2.

3) The parameters of the Pearsonian distribution (2.26) α and β are connected with its centre and variance:

$$a=\mathrm{E}\xi=\alpha/\beta,\ \sigma^2=\sigma^2(\xi)=\alpha/\beta^2,$$

see Exercise 7 to Chapter 3. Therefore, after calculating a and σ^2 by formula (4.14) and (4.17), we have

$$\alpha=a^2/\sigma^2,\ \beta=a/\sigma^2.$$

If the density of distribution is known to depend on l parameters α_1, α_2, …, α_l, then, expressing the first l moments of the distribution through them, we can, generally speaking, determine them; the moments themselves can be established by trials as stated above. We ought to remark, however, that, the higher the moment, the more data is needed for determining it more or less precisely. In practice therefore we often restrict our calculations to two unknown parameters (and two moments).

4.5. Exercises

1) Wrong connections were each minute registered at a telephone exchange during an hour. Here are the results [the author provided the number of those connections for each minute of the hour]. Determine the centre and variance of the distribution and check whether the main condition for the appearance of the Poisson distribution, $E\xi = \sigma^2 = a$, is fulfilled. Determine that distribution and compare the registered data with a table of that distribution. *Answer*:

The mean number of wrong connections was $\bar{x} = 2$. The condition for the Poisson distribution is $s_n^2 \approx 2.1 = \bar{x}$. The Poisson distribution is here

$$P(\xi = m) = \frac{2^m}{m!}e^{-2} \ (a = 2).$$

2) [The author provides a table of the deviations of the sizes of a hundred manufactured articles from the nominal size.] Determine the centre and variation of the distribution and construct the appropriate normal distribution. Compare the data with a table of that law. *Answer*.

Mean deviation $\bar{x} = 0.4$; the sample variance $s_n^2 = 2.57$. The appropriate normal distribution has $\sigma = 1.6$. Each deviation x should be considered as the mean for the appropriate interval. Thus, deviations $x = -3, -2, -1$, have frequencies 3, 10 and 15 and frequency 10 is attributed to the interval $-2.5 < x < -1.5$. The distribution function of the normal law is therefore taken as

$$F(x + 0.5) = \frac{1}{2}+\frac{1}{2}\Phi[\frac{x+0.5-a}{\sigma}]$$

where Φ is determined by formula (2.22).

3) Determine the Pearsonian distribution (2.26) for the following data $x = 0, 1, 2, 3, 4, 5, 6$ and frequency $m = 1, 33, 41, 18, 5, 1, 1$ (sum = 100). Check whether the main condition for the appearance of that distribution, $C_s = 2C_v$, see Exercise 10 in Chapter 3, is fulfilled. *Answer*.

$\bar{x} = 2.0$, $s_n^2 = 1.0$. For distribution (2.26) $a = \alpha/\beta = 2$, $\sigma^2 = \alpha/\beta^2 = 1$, so $\alpha = 4$ and $\beta = 2$. The density of the Pearsonian distribution is

$$\varphi(x) = (16/3!)x^3e^{-2x}, \text{ and } C_v = \sigma/a = 0.5,\ C_s = 1.09 \approx 2C_v.$$

Chapter 5. Limiting Theorems and Estimation [of the Precision] of the Means

The distribution of the probabilities of the relative frequency and some other means tend to the normal law which is the decisive circumstance influencing their estimation. For proving the appropriate limiting theorems Liapunov developed a very powerful method of characteristic functions which allowed him to prove the so-called central limit theorem. Before discussing limiting theorems we provide necessary information about those functions.

5.1. Notion of characteristic functions

The characteristic function of random variable ξ is the expectation

$$f(u) = \mathrm{E}e^{iu\xi}. \qquad (5.1)$$

Here the parameter u is a real number. For a discrete random variable

$$f(u) = \sum e^{iux_k} p_k \qquad (5.2)$$

where p_k is the probability of the value x_k and the sum covers all the values x_k of ξ. For a continuous random variable

$$f(u) = \int_{-\infty}^{\infty} e^{iux}\varphi(x)dx \qquad (5.3)$$

where $\varphi(x)$ is the density of the distribution of ξ. The integral always converges absolutely since $|e^{iux}\varphi(x)| = \varphi(x)$ and

$$|f(u)| \le \int_{-\infty}^{\infty} \varphi(x)dx = 1.$$

5.1.1. *Main properties of characteristic functions*

1) They uniquely determine the distribution of probabilities of random variables. It is possible to indicate the general expression of the distribution function through the characteristic function (Gnedenko 1954, Chapter 7). There also the reader will find the proofs of the properties of characteristic functions. For those acquainted with the Fourier integral I note that the integral (5.3) is the Fourier transform of density $\varphi(x)$. If two random variables have identical characteristic functions, they also have identical distributions of probabilities.

2) If the characteristic function $f(u)$ of a *continuous* random variable ξ is the limit of a sequence of characteristic functions $f_n(u)$ of any random variables ξ_n ($n = 1, 2, \ldots$) the distribution function $F(x) = P(\xi < x)$ is the limit of functions $F_n(x) = P(\xi_n < x)$. It follows that, as $n \to \infty$,

$\lim f_n(u) = f(u)$ leads to $\lim F_n(x) = F(x)$ for all x.

Gnedenko provides more general theorems of this kind.

This property is important since in many cases the passage to the limit for a sequence of characteristic functions is easier than for a sequence of distribution functions. The proof of limiting theorems through characteristic functions is then shorter and simpler. The properties stated above we provide without proof.

3) The characteristic function of a sum of independent random variables is the product of the characteristic functions of the terms of that sum. Suppose we have independent random variables ξ and η with characteristic functions $f_\xi(u)$ and $f_\eta(u)$. The random variables $e^{iu\xi}$ and $e^{iu\eta}$ will also be independent and the characteristic function $f_{\xi+\eta}(u)$ of the sum (ξ + η) can be calculated by the multiplication theorem for expectations

$$f_{\xi+\eta}(u) = \mathrm{E}e^{iu(\xi+\eta)} = \mathrm{E}e^{iu\xi}e^{iu\eta} = \mathrm{E}e^{iu\xi}\mathrm{E}e^{iu\eta}, f_{\xi+\eta}(u) = f_\xi(u)f_\eta(u). \quad (5.4)$$

The calculation of a characteristic function of a sum of independent random variables is therefore easier than the determination of the corresponding distribution of probabilities (which is reduced to the convolution of the densities of the distributions of the terms, see § 2.4.4).

4) When passing from a random variable ξ to its linear function $\eta = A + B\xi$, the characteristic function becomes

$$f_\eta(u) = e^{iAu}f_\xi(Bu). \quad (5.5)$$

This formula can be checked at once:

$$\mathrm{E}e^{iu\eta} = \mathrm{E}e^{iu(A+B\xi)} = e^{iAu}\mathrm{E}e^{iBu\xi}.$$

5.1.2. *Examples*

1) For random variable λ with values 1, 0 and probabilities *p* and *q* (§ 2.2) the characteristic function according to formula (5.2) is

$$f_\lambda(u) = e^{iu1}p + e^{iu0}q = pe^{iu} + q.$$

2) The frequency of a random event [in *n* trials] is the sum

$$\mu_n = \lambda_1 + \lambda_2 + \ldots + \lambda_n$$

with independent λ_k having distributions as in the previous example. By Property 3 the characteristic function of μ_n is

$$f_{\mu_n}(u) = f_{\lambda_1}(u)f_{\lambda_2}(u)...f_{\lambda_n}(u) = (pe^{iu} + q)^n. \quad (5.6)$$

3) The relative frequency of the same event is $w_n = \mu_n/n$. By formulas (5.5) and (5.6)

$$f_{w_n}(u) = f_{\mu_n}(u/n) = (pe^{iu/n} + q)^n.$$

4) A random variable uniformly distributed on (– *a, a*) has density

$\varphi(x) = 1/2a$, $-a < x < a$ and 0 otherwise.

By formula (5.3)

$$f(u) = \frac{1}{2a}\int_{-a}^{a} e^{iux}dx = \frac{e^{iua} - e^{-iua}}{2aiu} = \frac{\sin au}{au}. \quad (5.7)$$

5) Random variable ξ_0 has the simplest normal distribution (2.21). Its characteristic function is

$$f_0(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{iux}e^{-x^2/2}dx = \exp(-\frac{u^2}{2}). \qquad (5.8)$$

I am omitting the calculation of the integral. If random variable ξ has the general normal distribution (2.25), so that $\xi = a + \sigma\xi_0$, its characteristic function, according to formulas (5.8) and (5.5), is

$$f(u) = e^{iau}f_0(\sigma u) = e^{iau}\exp(-\frac{\sigma^2u^2}{2}). \qquad (5.9)$$

It follows that if mutually independent random variables $\xi_1, \xi_2, \ldots, \xi_n$ have normal distributions with parameters $a_1, a_2, \ldots, a_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, their sum has normal distribution with centre $a = a_1 + a_2 + \ldots + a_n$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$. Indeed, since

$$f_k(u) = \exp(ia_ku)\exp(-\sigma_k^2u^2/2), \; k = 1, 2, \ldots, n,$$
$$f(u) = f_1(u)f_2(u)\ldots f_n(u) = \exp(iau)\exp(-\sigma^2u^2/2).$$

5.1.3. *Connection between the characteristic function and the moments of distribution.* Since the characteristic function $f(u)$ uniquely determines the distribution of the probability of the appropriate random variable ξ, all the moments of the distribution can be expressed through it. Formally differentiate the equality

$$f(u) = \mathrm{E}e^{iu\xi}$$

with respect to u (after introducing either a sum or an integral):

$$f'(u) = \mathrm{E}i\xi e^{iu\xi}, f''(u) = \mathrm{E}(i\xi)^2e^{iu\xi}, \ldots, f^{(k)}(u) = \mathrm{E}(i\xi)^ke^{iu\xi}.$$

It can be shown that this is admissible if ξ has moments up to the k-th inclusive. Now, taking $u = 0$ we will have the connection sought

$$f(0) = \mathrm{E}1 = 1, f'(0) = i\mathrm{E}\xi, f''(0) = -\mathrm{E}\xi^2, \ldots, f^{(k)}(0) = i^k\mathrm{E}\xi^k.$$

In applications of probability theory derivations of $\psi(u) = \ln f(u)$ are sometimes needed. The number $i^k\psi^{(k)}(0)$ is called the k-th cumulant of ξ. It is easy to check that

$$i\psi'(0) = -\mathrm{E}\xi, \; i^2\psi'(0) = \sigma^2(\xi).$$

Cumulants are very important for calculating sums of independent random variables: their cumulants are then summed up as well.

5.2. The De Moivre – Laplace limiting theorem.

Estimating relative frequencies

We are now considering the limiting distribution of the relative frequency of a random event as the number of trials unboundedly increases. For sampling with replacement the distribution of the relative frequency w_n is binomial (§ 2.2):

$$P(w_n = \frac{m}{n}) = C_n^m p^m q^{n-m},\ m = 0, 1, 2, \ldots, n. \qquad (5.10)$$

Here n is the number of (independent) trials, p, the probability of the appearance of the studied event in each trial. For a large n calculations by formula (5.10) become very difficult. The essence of the appearing difficulties will be even clearer when noting that practically we are interested in the probability not of the equality $w_n = m/n$ but rather of the inequality $|w_n - p| < \varepsilon$ equal to the sum of the expressions in the right side of (5.10) covering the values of m for which $|m/n - p| < \varepsilon$ or the values of m satisfying inequalities $np - n\varepsilon < m < np + n\varepsilon$, see § 4.2.

Here is an example of such a difficulty. Required is the probability that after n = 10,000 trials the relative frequency of the event will not deviate from its probability $p = 0.2$ more than by $\varepsilon = 0.01$. Here np = 2000, $q = 0.8$. We have to sum up more than 200 terms such as

$$\frac{10{,}000!}{m!(10{,}000-m)!}0.2^m 0.8^{10{,}000-m},\ np - n\varepsilon = 1900 < m < np + n\varepsilon = 2100.$$

It was understood long ago that for an approximate calculation of probabilities the binomial distribution should be replaced by some continuous limiting law. Continuous, since then the problem is reduced to calculating an integral which is usually much easier than calculating sums in case of discrete distributions.

De Moivre solved it in 1730 for $p = q = 1/2$ [in 1733 for the general case], then Laplace in 1783 followed suit[10]. It occurred that a binomial distribution has a limiting law ($n \to \infty$), the normal distribution. We will first norm w_n. A [centred and] normed random variable ξ is

$$\xi_0 = (\xi - E\xi)/\sigma(\xi).$$

Denote the normed relative frequency by τ_n:

$$\tau_n = \frac{w_n - Ew_n}{\sigma(w_n)} = \frac{w_n - p}{\sqrt{pq/n}}. \qquad (5.11)$$

5.2.1. *The De Moivre – Laplace theorem.* As the number of trials unboundedly increases, the simplest normal law becomes the limiting distribution of probabilities of the normed relative frequency of a random event

$$\lim P(|\tau_n| < t) = \Phi(t),\ n \to \infty \qquad (5.12)$$

where $\Phi(t)$ is the integral (2.22).

This theorem is a particular case of a more general proposition (§ 5.3). Here, we only indicate the manner of its application.

5.2.2. *Application of the De Moivre – Laplace theorem to the estimation of relative frequencies*. This theorem ensures the estimation of probabilities of the inequality $|w_n - p| < \varepsilon$ for sufficiently large values of n (and values of p not too near to 0 or 1). So let us choose such a large value of n that the approximate equality

$$P(|\tau_n| < t) \approx \Phi(t) \qquad (5.13)$$

will be obeyed with a satisfactory precision.

Then, since the inequalities

$$|w_n - p| < \varepsilon \text{ and } |\tau_n| = \frac{|w_n - p|}{\sqrt{pq/n}} < \frac{\varepsilon\sqrt{n}}{\sqrt{pq}} = t \qquad (5.14)$$

are equivalent, the probability of the former is approximately equal to $\Phi(t)$.

5.2.3. *Notion of interval estimation*. Return to the example in § 5.2: $p = 0.2$, $q = 0.8$, $n = 10{,}000$ and $\varepsilon = 0.01$. Since

$$t = \frac{0.01\sqrt{10{,}000}}{\sqrt{0.2\cdot 0.8}} = 2.5$$

probability $P(|w_n - 0.2| < 0.01) \approx \Phi(2.5) = 0.988$.

Since n is rather large this approximate formula ensures the third decimal place. Precise estimates are provided in special papers (Bernstein, Feller et al). For n of the order of several hundred (but np and nq nevertheless considerably larger than 1) a somewhat more precise formula with integer k is

$$P(|w_n - p| \le k/n) \approx \Phi[\frac{k+1/2}{\sqrt{npq}}].$$

I illustrate the precision of this new formula by two examples. […]

If $P \approx 0.988$ is considered sufficiently near to unity, we are practically certain that $P(|w_n - p| < \varepsilon) \approx \Phi(t)$ will be fulfilled here. That value, 0.988, is then called the confidence probability of the estimate $|w_n - p| < \varepsilon$. Confidence probability is assigned beforehand in accordance with the stipulated boundary of very low probabilities (§ 4.1). Thus, if we decide to neglect the possibility of the appearance of an event having probability 0.001, the confidence probability will be 0.999.

Knowing P we calculate t by issuing from equation $\Phi(t) = P$ and an appropriate table of the normal distribution. Thus, for $P = 0.999$, we find $t = 3.29$ and with confidence probability P we have

$$|w_n - p| < \varepsilon,\ \varepsilon = t(p)\sqrt{pq/n}. \qquad (5.15)$$

With an assigned confidence probability P the relative frequency w_n will be contained within the confidence interval $(p - \varepsilon, p + \varepsilon)$ with ε provided in formula (5.14).

For our example with confidence probability $P = 0.999$,

$$|w_n - 0.2| < 3.29\sqrt{\frac{0.2 \cdot 0.8}{10{,}000}} = 0.0132$$

and the relative frequency w_n is contained within confidence interval (0.1868, 0.2132). For frequency $\mu_n = nw_n$ that interval will be $n =$ 10,000 times wider: $1868 < \mu_n < 2132$. Confidence intervals are practically useful not only by allowing us to foresee the boundaries of (relative) frequencies. If the trials were really carried out and the actual frequency was beyond the confidence interval, we ought to doubt the results of calculating the confidence probability of the studied event.

This is important for, say, regulating mass production. Suppose that substandard manufactured articles ought to comprise no more than 1% of the total. A random sampling inspection with replacement is needed. However, if the sample size (n) is very small as compared with the total number of the articles, the formulas above will be sufficiently precise even without replacing the selected articles. The frequency μ_n $= nw_n$ of substandard articles should be contained within interval

$$n(p - \varepsilon) < \mu_n < n(p + \varepsilon).$$

5.3. Confidence estimation of means. Notion of the Liapunov central limit theorem

The normal limiting law in the De Moivre – Laplace theorem is not connected with some specific properties of the binomial distribution. It is only occasioned by relative frequency w_n being an arithmetic mean of independent random variables $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$ That theorem can be directly generalized on arithmetic means of any sequence of independent identically distributed random variables (if the centre and the variance of their distribution are finite).

Suppose that $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ is such a sequence of variables with centre and variance of their distribution being $E\xi_k = a$ and $E(\xi_k - a)^2 =$ σ^2, $k = 1, 2, \ldots$ Compile

$$\overline{\xi}_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n}$$

and introduce normed [and centred] means

$$\tau_n = \frac{\overline{\xi}_n - E\overline{\xi}_n}{\sigma(\overline{\xi}_n)}. \qquad (5.16)$$

The variables ξ_k are independent and therefore (cf. Chapter 3)

$$\mathrm{E}\overline{\xi}_n = a,\ \sigma(\overline{\xi}_n) = \sigma/\sqrt{n},$$

$$\tau_n = \frac{\overline{\xi}_n - a}{\sigma/\sqrt{n}} = \frac{\xi_1 + \xi_2 + \ldots + \xi_n - na}{\sigma\sqrt{n}}.$$

Theorem. The limiting ($n \to \infty$) *distribution of the normed means* (5.16) *is the normal law*

$$\lim P(|\tau_n| < t) = \Phi(t) \qquad (5.17)$$

where Φ *is the integral* (2.22).

We will prove it only for continuous random variables by applying characteristic functions. […]

This theorem allows us to derive interval estimates of the means; that is, to estimate ε in the inequality $|\overline{\xi}_n - a| < \varepsilon$ with an assigned confidence probability P. Replace that inequality by an equivalent and therefore equally probable inequality

$$\frac{|\overline{\xi}_n - a|}{\sigma/\sqrt{n}} < \frac{\varepsilon\sqrt{n}}{\sigma} \text{ or } |\tau_n| < t = \frac{\varepsilon\sqrt{n}}{\sigma}.$$

According to (5.13), the probability of $|\tau_n| < t$ will also be approximately equal to $\Phi(t)$, $t = \varepsilon\sqrt{n}/\sigma$. Assuming a definite probability P sufficiently near to 1, we can find $t = t(P)$ satisfying equation $\Phi(t) = P$ by a table of the normal law and thus derive an interval estimate of the mean $\overline{\xi}_n$:

$$|\overline{\xi}_n - a| < \varepsilon = t(P)\sigma/\sqrt{n} \text{ with confidence probability } P. \quad (5.18)$$

5.3.1. *Deviations of the experimental mean from the expectation.* We are now interested in a *single* random variable ξ with centre a and variance σ^2. Suppose that a sufficiently large number n of independent trials were made to find its particular values. Whatever are those values $x_1, x_2, \ldots, x_n$, we may state with probability P that the mean $\overline{x}$ will obey the inequality

$$|\overline{x} - a| < t(P)\sigma/\sqrt{n} \qquad (5.19)$$

or that it will be contained in a confidence interval $(a - \varepsilon, a + \varepsilon)$, $\varepsilon = t(P)\sigma/\sqrt{n}$. This statement follows from estimate (5.18) if only we connect random variable ξ_k having the same distribution as ξ with each k-th trial. The particular values of ξ_k and $\overline{\xi}_n$ will be x_k and $\overline{x}$. The independence of ξ_k follows from the supposed independence of the trials.

5.3.2. *Deviations of the sample mean from the general mean.* Consider the possible values of the indication of each element of the sample as a random variable with distribution (3.2). Then the sample mean $\bar{x}$ will be the arithmetic mean of these values. It will therefore satisfy inequality (5.18) where a is the general mean, the centre of distribution (3.2), and

$$\sigma = \sqrt{(x_1 - a)^2 M_1/N + (x_2 - a)^2 M_2/N + ... + (x_v - a)^2 M_v/N}.$$

In other words, we can expect with probability P that the sample mean deviates from the general mean a not more than by $t(P)\sigma/\sqrt{n}$. Interval estimation of the mean can be applied for checking and regulating manufacturing when some parameter (a size, for example) ought to be kept within definite boundaries. If a sample inspection shows that some mean value is contained beyond the interval ($a - \varepsilon$, $a + \varepsilon$), $\varepsilon = t(P)\sigma/\sqrt{n}$, it will be necessary to check whether the conditions of manufacturing were not violated. The development of such principles led to the creation of special methods of statistical inspection.

5.3.3. *Notion of the Liapunov central limit theorem (CLT).* Above, we have established that the normal law is the limiting distribution for normed means (for the normed sums of identically distributed terms). The CLT establishes the general conditions for the normal distribution to be the limiting law of normed sums of mutually independent random summands. In a general form this problem was first formulated in Chebyshev's researches, but the conditions which he found were rather restrictive[11]. In 1900, Liapunov proved the CLT under very general conditions, proved the sufficiency of two conditions:

1) All the random summands have finite absolute central moments of the third order

$$\mathrm{E}|\xi_k - a_k|^3, \ a_k = \mathrm{E}\xi_k, \ k = 1, 2, \ldots$$

$$\textbf{2)} \sum_{k=1}^{n} |\xi_k - a_k|^3 \div \{\sum_{k=1}^{n} \sigma^2(\xi_k)\}^{3/2} \to 0 \text{ as } n \to \infty. \qquad (5.20)$$

This second condition is satisfied for identically distributed summands since then

$$n\mathrm{E}|\xi - a|^3 \div (n\sigma^2)^{3/2} = \frac{1}{\sqrt{n}} \frac{\mathrm{E}|\xi - a|^3}{\sigma^3}.$$

The Liapunov conditions thus mean an *utmost neglect* of separate summands of a sum, a uniformly small influence of each of them on the sum. This is clearer seen in the somewhat more general Lindeberg conditions[12] which require a uniform smallness of the probabilities of large deviations $|\xi_k - a_k|$ as compared with the variance of the sum $\sigma^2(\xi_1 + \xi_2 + \ldots + \xi_n)$. Roughly speaking, there should be no summands whose possible deviations dominate those of all the other ones.

Liapunov's CLT explained the prevalence of the normal law in nature and technology: the scattering of the studied magnitude is caused by a very large number of random causes whose separate influence is negligible. On the other hand, it established the precise conditions of the CLT and thus strictly defined the applicability of the normal law.

The importance of the law of large numbers and the CLT for the entire probability theory and its applications led to their numerous and most various specifications and generalizations. In particular, we note the study of dependent random variables originated by Markov and continued by Bernstein and Slutsky. The Chebyshev theorem, for example, proved to be also valid for sequences of dependent random variables with restricted variances if only that dependence rapidly lessened with the distance between the terms of those sequences. It suffices that the correlation coefficient (§ 7.4)

$$r(\xi_i, \xi_k) \to 0 \text{ as } |i - k| \to \infty.$$

Similar conditions concerning the dependence can ensure the generalization of the CLT on dependent variables.

5.4. Exercises

1) Directly prove the De Moivre – Laplace theorem by applying the characteristic function of the frequency μ_n (5.6).

Indication: Derive the characteristic function $f_n(u)$ of the normed frequency

$$\tau_n = \frac{\mu_n - np}{\sqrt{npq}} \qquad (5.11)$$

by formulas (5.6), (5.5) and (5.11); then, after expanding the exponents

$$\exp[\frac{iu\sqrt{q/p}}{\sqrt{n}}] \text{ and } \exp[-\frac{iu\sqrt{p/q}}{\sqrt{n}}]$$

transform $f_n(u)$ into

$$[1 - \frac{u^2}{2n} - i\frac{u^3}{3!n\sqrt{n}} - \frac{q-p}{\sqrt{pq}} + ...]^n.$$

2) Estimate the relative frequency w_n assuming $p = 0.01$, $n = 1000$, and $P = 0.99$. *Answer*:

$$|w_n - p| < 2.576\sqrt{0.0099/1000} = 0.0081, \; 0.0019 < w_n < 0.0181.$$

With confidence probability $P = 0.99$ we may expect that in a thousand trials the studied random event will occur 2 – 18 times ($2 \le \mu_n = nw_n \le 18$).

3) A coin is tossed 12,000 times and heads appeared in 6019 tosses. Does this agree with the supposed probability of heads being 1/2? *Answer*:

$$P(|w_n - 1/2| \geq 19/12{,}000) \approx 1 - \Phi[\frac{19\sqrt{12{,}000}}{12{,}000\sqrt{0.5\cdot 0.5}}] = 0.738.$$

This probability is not low and doubts about the formulated hypothesis are unfounded.

4) The distribution of some indication in a batch of 5000 articles is:

values: 3.40(0.05)3.75
frequencies M: 150, 380, 1320, 1530, 970, 470, 100, 80

A sample contains 100 articles. Estimate the sample mean with P = 0.99. *Answer*:

General mean $a = 3.55$, $\sigma = 0.05\sqrt{1.844} = 0.068$,

$$|\bar{x} - a| < 2.576\frac{0.068}{\sqrt{100}} = 0.0175,\ 3.5325 < \bar{x} < 3.5675.$$

5) A random sample of a batch of a hundred articles resulted in

values: same as in Exercise 4; frequencies m: 3, 5, 12, 28, 28, 14, 8, 2

Suppose that σ is also the same as in the previous Exercise. May we decide that the mean value of the indication is the same just as well? *Answer*:

$$\bar{x} = 3.55 + 0.05\cdot 57/100 = 3.55 + 0.0285.$$

$$P(|\bar{x} - a| \geq 0.0285) \approx 1 - \Phi[\frac{0.0285\sqrt{100}}{0.068}] < 0.00003.$$

This is a very low probability and we cannot think that the new batch has mean value $a = 3.55$.

6) Prove that the Poisson distribution can be considered the limiting case of the binomial law ($n \to \infty$, $p \to 0$, $np = a$).

Indication: Replace p by a/n and pass to the limit in formula

$$C_n^m p^m q^{n-m} = C_n^m a^m / n^m (1 - a/n)^{n-m}.$$

7) A large number n of terms rounded off to 10^{-m} is summed up. The error of the approximation is supposed to be a random variable ξ uniformly distributed on $(-\,0.5\cdot 10^{-m}, 0.5\cdot 10^{-m})$. Show that the absolute error of the sum will not exceed $0.5\cdot 10^{-m}\sqrt{3n}$ with $P = 0.997$.

Indication. Suppose that the errors of the summands are independent and identically distributed and that having a sufficiently large n their sum is distributed near-normally with centre 0 and mean square deviation

$$\sigma = \sigma(\xi)\sqrt{n} = \frac{10^{-m}\sqrt{n}}{2\sqrt{3}},$$

see Exercise 2 in § 3.3.2.

Chapter 6. Application of Probability Theory to Mathematical Treatment of Observations

6.1. Random observational errors, their distribution

An error of measurement is the difference x – a between the measurement and the true value a of the measured magnitude.

Each measurement is corrupted by errors. The results of measurements of the same magnitude even repeated under identical conditions usually differ and the measurements themselves and any result of their treatment only provide approximate rather than precise values of *a*. From all such approximations we have to select in some sense the best one. Then, we ought to estimate the precision of the obtained approximation, i. e., to establish the boundaries which with a given probability the deviation of the true value from that approximation will never exceed.

The applicability of the probability theory to these problems is based on the fact that the possible result of a measurement is a random variable with a definite distribution of probabilities. Let us establish the type of this distribution in case of *direct measurements* (when their results are directly read on a scale). We assume that *the results of measurements are not affected by systematic errors*. Those errors are caused by an invariably acting cause; their magnitudes are either identical in all measurements or vary according to a known law. They can therefore be eliminated by regulating the measuring device or appropriately correcting the results of measurements[13].

After that these results will still be corrupted by unavoidable errors which are impossible to get rid of, by *random errors*. They are caused by numerous and hardly perceptible causes each of which only leads to small fluctuations of the results. Each of those causes generates its so-called elementary error and the resulting error is obviously their sum. If the number of elementary errors is very large and the contribution of each of them is very small (which is the essence of *the hypothesis of elementary errors*), then, according to the CLT, the resulting error should more or less obey the normal law. The decisive argument in favour of that law is its confirmation by numerous experiments and observations. The theory of errors therefore assumes as the main axiom that the random error τ of a direct measurement obeys the normal law[14].

Then, *blunders* occur when the stipulated conditions of measurement are violated or the result of a measurement is wrongly recorded. The thus corrupted results ought to be rejected at once.

Since the results of measurements are read on scales, random errors are always expressed by some [rational] numbers, but it is more convenient to consider them continuous. And for the same reason we assume that these errors take any value on the numerical axis. This assumption sometimes contradicts the essence of a problem, but does not influence the conclusions since the probability of τ exceeding definite boundaries is very low.

Taking into account the usual symmetry of random errors [see below], it is also assumed that the centre of their distribution is zero so that the density of their distribution is the normal law with parameters 0 and σ^2; $\sigma = \sqrt{E\tau^2}$ is called mean square error of measurement or standard. It characterizes the precision of measurement (or of the measuring device).

The possible result of a measurement, ξ, and τ are connected by a simple equality

$$\xi = a + \tau.$$

Since $E\tau = 0$, $E\xi = a$, which is the condition of unbiasedness practically connected with the absence of systematic errors. And so, ξ obeys the normal law with parameters a and σ^2.

6.2. Solution of the two main problems of the error theory

Suppose that $x_1, x_2, \ldots, x_n$ are the results of direct measurements of some constant magnitude a. We assume that the possible results of measurements $\xi_1, \xi_2, \ldots, \xi_n$ obey the normal law with an identical centre

$$E\xi_k = a,\ k = 1, 2, \ldots, n \qquad (6.1)$$

(unbiasedness) and identical variances

$$E(\xi_k - a)^2 = \sigma^2,\ k = 1, 2, \ldots, n \qquad (6.2)$$

(measurements of equal precision).

It is advisable to assume the arithmetic mean $\overline{x}$ as the approximate value of a (§ 4.4.1). Now, however, we have to estimate the precision of that approximation. Random variables ξ_k are independent and normally distributed with parameters a and σ^2 so that (§§ 5.2 and 3.3.1) the mean $\overline{\xi}$ is also normally distributed with parameters a and σ^2/n. Therefore

$$P(|\overline{\xi} - a| < \varepsilon) = \Phi(t),\ t = \varepsilon/\sigma(\overline{\xi}) = \varepsilon\sqrt{n}/\sigma. \qquad (6.3)$$

Here, $\Phi(t)$ is the integral (2.22).

The probability P is usually assigned beforehand and is near 1 (for example, 0.999). Therefore, t is derived from equation $\Phi(t) = P$ by means of a table of the normal law. Thus, if $P = 0.999$, $t = 3.291$. And so we obtain

$$|\overline{\xi} - a| < \varepsilon) = t\sigma/\sqrt{n}.$$

Replacing $\overline{\xi}$ by $\overline{x}$ we get the so-called classical estimate

$$|\overline{x} - a| < t\sigma/\sqrt{n},\ \overline{x} - t\sigma/\sqrt{n} < a < \overline{x} + t\sigma/\sqrt{n}. \qquad (6.4)$$

The probability of those inequalities is the assigned number $P = \Phi(t)$.

This estimate has as essential defect: it assumes that σ^2 is known. When replacing this variance by its approximate value (§ 4.4.1),

$$\sigma^2 \approx s_n^2 = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n-1} \qquad (6.5)$$

the confidence probability of (6.4) decreases. However, it occurs that a proper estimate of the precision sought is possible when issuing not from $\bar{\xi} - a$, but from another random variable,

$$\varsigma = \frac{\bar{\xi} - a}{\sqrt{n(n-1)}\sqrt{\sum_{k=1}^{n}(\xi_k - \bar{\xi})^2}}, \; n \geq 2.$$

If all the ξ_k are independent and normally distributed with centre a, ς we get the so-called Student distribution with density

$$S(t,\, n) = B_n(1 + \frac{t^2}{n-1})^{-n/2}, \; B_n = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma[(n-1)/2]}.$$

Number n is supposed fixed.

The probability of $|\varsigma| < t$ is therefore

$$P = \int_{-t}^{t} S(t, n)dt. \qquad (6.6)$$

Having a table of that integral and assuming a given P we can calculate $t = t(P, n)$ and, taking (6.6) into account, we have the estimate sought

$$\frac{|\bar{x} - a|}{s_n/\sqrt{n}} < t = t(P, n), \; \bar{x} - ts_n/\sqrt{n} < a < \bar{x} + ts_n/\sqrt{n} \qquad (6.7)$$

[…]

6.2.1. *Calculation of the means*. For applying (6.7) we ought to calculate $\bar{x}$ and s_n which can be essentially simplified by an appropriate linear transformation of the results of measurement

$$x_k = c + hu_k,\; u_k = (x_k - c)/h,\; k = 1, 2, \ldots, n. \qquad (6.8)$$

The chosen c is some mean between the extreme values of x_k, and h can always be selected in such a way that u_k will be integers since the results of measurements are rational numbers (§ 6.1). The necessary formulas are

$$\bar{x} = c + h\bar{u}, \qquad (6.9)$$

$$\sum(x_k - \bar{x})^2 = h^2(\sum u_n^2 - n\bar{u}^2) = z,\ s_n = \sqrt{z/(n-1)}. \quad (6.10)$$

Example. Here are the results of the 20 first measurements of the elementary electronic charge made by Millikan. […] Choose $c = 4.780$ and $h = 0.001$ [in appropriate units]. The sums of their deviations and of their squares are – 29 and 1871. By formulas (6.9) and (6.10) $\bar{u} = -29/20 = -1.45$,

$$\bar{x} = 4.780 - 0.00145 = 4.77855,\ n\bar{u}^2 = 20(29/20)^2 = 42.0,$$

$$s_n = 0.001\sqrt{\frac{1871-42}{19}} = 0.00981.$$

The true charge can be assumed as $e = 4.7786$. Let the confidence probability be $P = 0.99$. Then for $n = 20$ we find $t = 2.861$ and we may maintain that the true value of the charge is contained between

$$\bar{x} - ts_n\sqrt{n} = 4.77855 - 2.861 \cdot 0.00981/\sqrt{20} = 4.7722 \text{ and}$$

$$\bar{x} + ts_n\sqrt{n} = 4.77855 + 2.861 \cdot 0.00981/\sqrt{20} = 4.7848$$

If however $P = 0.999$, $t = 3.883$ and $4.7700 < e < 4.7870$.

6.2.2. *Estimating the precision of the device (of measurements).* Precision is measured by σ; its approximate value is the sample standard s_n [see (6.5)]. For estimating the approximate equality $\sigma \approx s_n$ we may apply the distribution of probabilities of the random variable

$$\chi = \frac{1}{\sigma}\sqrt{\sum(\xi_k - \bar{\xi})^2}$$

[Cramér 1946/1948, § 18.1] which depends on n but not on a or σ. If all the ξ_k, $k = 1, 2, \ldots, n$ are independent and have an identical normal distribution with parameters a and σ^2 the density of χ will be

$$R(t, n) = A_n t^{n-2}\exp(-t^2/2),\ n \geq 3,\ t \geq 0,\ A_n = 1 \div 2^{n-3/2}\Gamma[(n-1)/2],$$

$$P(t_1 < \chi < t_2) = \int_{t_1}^{t_2} R(t,n)dt. \quad (6.11)$$

We can now determine the probability of

$$s_n - \varepsilon < \sigma < s_n + \varepsilon \text{ or } s_n(1 - q) < \sigma < s_n(1 + q) \quad (6.12)$$

where $q = \varepsilon/s_n$ is the relative error. Indeed, inequalities

$$(1-q)\sqrt{\frac{\sum(\xi_k - \bar{\xi}\ )^2}{n-1}} < \sigma < (1+q)\sqrt{\frac{\sum(\xi_k - \bar{\xi}\ )^2}{n-1}} \quad (6.13)$$

can be transformed into identical and therefore equally probable inequalities

$$\frac{\sqrt{n-1}}{1+q} < \frac{1}{\sigma}\sqrt{\sum(\xi_k - \bar{\xi})^2} < \frac{\sqrt{n-1}}{1-q},\ q < 1.$$

They are of the type $t_1 < \chi < t_2$ and their probability is therefore equal to the integral (6.11) with

$$t_1 = \frac{\sqrt{n-1}}{1+q},\ t_2 = \frac{\sqrt{n-1}}{1-q}.$$

After assigning a definite P we determine $q = q(P, n)$ from integral (6.11) taken with these t_1 and t_2 and equating it to P. Then the inequalities (6.13) will have the assigned probability P. Replacing ξ_k by experimental x_k, we will obtain estimate (6.12) with that assigned probability.

Note. If $q > 1$, the inequalities (6.13) will become

$$0 < \sigma < (1+q)\sqrt{\frac{\sum(\xi_k - \bar{\xi}\ )^2}{n-1}}$$

which are identical with

$$t_1 < \chi < \infty,\ t_1 = \frac{\sqrt{n-1}}{1+q}$$

and the probability P of those inequalities will be equal to the integral (6.11) taken over t_1 and ∞. Then $q = q(P, n)$ and estimate (6.13) becomes

$$0 < \sigma < s_n(1 + q).\ [\ldots]$$

Example. In the previous example we derived $s_n = 0.00981$. The precision of the Millikan measurements is therefore characterized by standard $\sigma \approx 0.00981$. Let us estimate this approximate equality assuming $P = 0.99$. We apply the table of $q = q(P, n)$ from Romanovsky (1947). For that value of P and $n = 20$, $q = 0.58$. We may therefore state that the standard error is contained within the interval

$$s_n(1 - q) = 0.00981\ (1 - 0.58) = 0.0041$$
$$s_n(1 + q) = 0.00981\ (1 + 0.58) = 0.00145$$

If however $P = 0.999$, $q = 0.88$ then $0.0012 < \sigma < 0.0185$. This interval can be shortened by an essentially larger number of measurements. For example, if $P = 0.99$ and 0.999, 350 and even 600 measurements are needed to derive the standard error of σ with relative precision of 10%.

6.2.3. *Simplified estimation. The three-sigma rule*[14]. The estimation discussed above demanded the study of special distributions (the Student, the χ distribution etc). In practice, estimation is often simplified, for example, by applying the three-sigma rule so that the error of the approximate equality $a \approx \bar{x}$ does not exceed three mean square errors of $\bar{x}$. So, if σ is known,

$$|a - \bar{x}| < 3\sigma(\bar{x}) = 3\sigma/\sqrt{n}$$

with confidence probability $P = \Phi(3) = 0.997$, see (6.4). The same rule is, however, applied when the unknown σ is replaced by s_n:

$$|a - \bar{x}| < 3s_n/\sqrt{n}. \qquad (6.14)$$

But then the confidence probability becomes considerably lower than 0.997 and decreases with n. Indeed, comparing estimates (6.14) and (6.4) we note that with $n = 14$ and 8 the former has $P < 0.99$ since $t(0.99, 14) = 3.01 > 3$ and the latter has $P = 0.98$.

Since calculating mean square errors is always simpler than studying the appropriate distributions, the same rule is again applied for estimating other characteristics of distributions. As an example, we describe now the estimation of the standard error of measurements. It can be shown that the mean square error of the sample standard is approximately

$$\sigma\sqrt{\frac{\sum(\xi_k - \bar{\xi}\,)^2}{n-1}} \approx \frac{s_n}{\sqrt{2(n-1)}}$$

so that the three-sigma rule becomes

$$|\sigma - s_n| < 3s_n/\sqrt{2(n-1)}\,. \qquad (6.15)$$

A comparison of it with the estimate (6.12) shows that even for $n = 45$ it has probability $P < 0.99$ since $q\,(0.99, 45) = 0.321 > 3/\sqrt{2(45-1)}$. For $n = 19$ and 7 the estimate (6.15) has $P = 0.98$ and lower than 0.95.

When estimating σ the confidence probability can be heightened not only by increasing n, but by measuring several magnitudes by the same device. Let $n_1, n_2, \ldots, n_m$ be the numbers of measurements of the first, the second, the n-th magnitude, and $s_1, s_2, \ldots, s_m$, the corresponding sample standards. Then, see for example Arley & Buch (1949), the three-sigma rule becomes

$$|\sigma - S| < 3S\sqrt{2(n-m)},\ S = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_\nu - 1)s_m^2}{n-m}},$$

$n = n_1 + n_2 + \ldots + n_m$. If $n - m = 200$ P reaches 0.995.

6.3. Exercises

1) Estimate the true value of the elementary electronic charge *e* for the 58 measurements made by Millikan […] choosing $P = 0.999$.
Answer:

The mean result of the measurements[15]

$$\bar{x} = 4.780 + 0.001 \cdot 0.81 = 4.78081 \approx e,$$

$$s_n/\sqrt{n} = 0.001\sqrt{\frac{13367-38}{57 \cdot 58}} = 0.00201, \quad |e - \bar{x}| < ts_n/\sqrt{n},$$

$$t = t(0.999, 58) = 3.470, \ \varepsilon = ts_n/\sqrt{n} = 0.00697, \ 4.7738 < e < 4.7878.$$

2) Same type of problem: $P = 0.99$ and the results of 100 measurements are

x_k: 3.18(0.02)3.28; frequency m: 4, 18, 33, 35, 9, 1

Indication: When calculating $\bar{x}$ and the sample variance s_n^2 take into account the frequencies. Denote $u = (x - x_0)/h$, $h = 0.02$, $x_0 = 3.22$.

$$\bar{x} = 3.22 + 0.02 \cdot 0.3 = 3.226 \approx a, \ s_n = 0.02\sqrt{(114-9)/99} = 0.0206 \approx \sigma.$$

Estimate of a: $t(0.99, 100) = 2.627$, $\varepsilon = ts_n/\sqrt{n} = 0.0054$,

$$|a - 3.226| < 0.0054.$$

Estimate of the standard σ of the random errors:

$$0.0206(1 - q) < \sigma < 0.0206(1 + q), \ q(0.99, 100) = 0.198,$$
$$0.0165 < \sigma < 0.0247.$$

3) Estimate by the three-sigma rule the precision of measurements of differing magnitudes made by the same device, of 10 of their series containing 15 measurements each. […]

Chapter 7. Linear Correlation

7.1. On different types of dependences

Functional connection between magnitudes is the simplest connection: a quite definite value of one corresponds to each value of the other. Examples: pressure and volume of a gas including connections between several arguments.

However, not all connections are functional: rainfall and crop yield; levels of accumulated snow and of later high water. Here, numerous possible values of a magnitude correspond to each value of another magnitude. The scattering of the former is caused by a large number of additional factors which we leave aside. Actually, most often we restrict our attention to studying the change of the mean characteristics of a magnitude caused by the change of the other one and the dependence of the calculated means on that other one will be functional.

Suppose that in an experiment each value of *x* led to several values of *y*. Its change with *x* can then be characterized by a broken line passing through each mean value of y corresponding to the appropriate x. Note that the dependence of x on y leads to another broken line sometimes essentially differing from the former line.

Definition. Two random variables, ξ and η, are correlatively dependent, if a definite distribution of probabilities of either of them corresponds to each value of the other. Such distributions are called *conditional*.

Here, we only discuss various means for conditional distributions of probabilities, and, in particular, the centres of those distributions.

7.2. Conditional expectations and their properties

The centre of a conditional distribution of η (its conditional expectation) when $\xi = x$ is defined as the sum of the products of the possible values of η by their conditional probabilities:

$$\mathrm{E}_x\eta = \sum yP(\eta = y|\xi = x). \tag{7.1}$$

Here, $P(\eta = y|\xi = x)$ is the conditional probability of the equality $\eta = y$ if $\xi = x$, and the sum covers all the values y of magnitude η. For continuous distributions that sum is replaced by the integral

$$\mathrm{E}_x\eta = \int_{-\infty}^{\infty} y\varphi_x(y)dy. \tag{7.2}$$

Here, $\varphi_x(y)$ is the density of the conditional distribution of probabilities of η if $\xi = x$. The conditional expectation $\mathrm{E}_x\eta$ is a function of x and is called regression function of η on ξ and denoted by $f(x)$: $f(x) = \mathrm{E}_x\eta$.

The equation $y = f(x)$ is the equation of regression of η on ξ and the corresponding line, the line of that regression. Regression of ξ on η is similarly defined as

$$\mathrm{E}_y\xi = \sum P(\xi = x|\eta = y) = g(y).$$

If the connection between ξ and η is not strictly functional, functions $f(x)$ and $g(y)$ are not mutually inverse, and the lines of regression of η on ξ and ξ on η do not coincide. If ξ and η are independent, then, for each x, $P(\eta = y|\xi = x) = P(\eta = y)$, therefore $\mathrm{E}_x\eta = \mathrm{E}\eta$ and formula (7.4) transforms into a simpler formula (3.9). The same is true in a more general case in which $f(x)$ is constant. Indeed, if $f(x) = b$, than, by formula (7.3) $\mathrm{E}\eta = \mathrm{E}b = b$, and, according to formula (7.5)

$$\mathrm{E}\xi\eta = \sum xbP(\xi = x) = b\mathrm{E}\xi = \mathrm{E}\xi\mathrm{E}\eta.$$

Formulas (7.3, 7.4 and 7.5) are below. The determination and study of regression functions is a main problem of correlation analysis. Important for linear correlation are formulas

$$\mathrm{E}\eta = \mathrm{E}f(\xi),\ \mathrm{E}\xi\eta = \mathrm{E}\xi f(\xi). \qquad (7.3),\ (7.4)$$

The latter can be considered as a generalization of the multiplication theorem for expectations on dependent random variables. Indeed, applying the general rule of multiplication of probabilities (1.13), we have

$$E\xi\eta = \sum xyP(\xi = x, \eta = y) = \sum\sum xyP(\xi = x)P(\eta = y|\xi = x).$$

The sums cover all possible values x and y of ξ and η respectively.

We have

$$E\xi\eta = \sum xP(\xi = x)\sum yP(\eta = y|\xi = x) = \sum xP(\xi = x)E_x\eta = \sum xf(x)P(\xi = x) \qquad (7.5)$$

which coincides with $E\xi f(\xi)$, see §§ 3.11, 3.12.

Formulas (7.3) and (7.4) are particular cases of a more general relation

$$Eu(\xi)\eta = Eu(\xi)f(\xi) \qquad (7.6)$$

where $u(\xi)$ is any function having $Eu(\xi)\eta$.

7.2.1. *Proof of formula (7.6).* We assume that the density $p(x, y)$ of the two-dimensional distribution of (ξ, η) is known. The probability of

$$x < \xi < x + dx \text{ and } y < \eta < y + dy \qquad (7.7a), (7.7b)$$

can be expressed as the probability of the product of those events. The general multiplication rule for probabilities leads to

$$p(x, y)dxdy = \psi_1(x)dx\varphi_x(y)dy,\ \psi_1(x) = \int_{-\infty}^{\infty} p(x, y)dy. \ (7.8)$$

Here, $\psi_1(x)$ is the density of the distribution of ξ and $\varphi_x(y)dy$ is the differential of the conditional probability of (7.7b) if $\xi = x$. Formula (7.8) indicates that

$$\varphi_x(y) = p(x, y)/\psi_1(x).$$

This can be substituted in formula (7.2) so that

$$f(x) = E_x\eta = \int_{-\infty}^{\infty} y\frac{p(x, y)}{\psi_1(x)}dy$$

and we obtain formula (7.6):

$$Eu(\xi)f(\xi) = \int_{-\infty}^{\infty}[u(x)f(x)]\psi_1(x)dx = \int_{-\infty}^{\infty} u(x)\psi_1(x)[\int_{-\infty}^{\infty} y\frac{p(x, y)}{\psi_1(x)}dy]dx = \int\int[u(x)y]p(x, y)dxdy = Eu(\xi)\eta.$$

Formula (7.6) shows that the mean value of any function $u(\xi)\eta + v(\xi)$ linear with respect to η does not change when η is replaced by the regression function of f on ξ. One more important property of the regression function follows: the mean square deviation of η from $f(\xi)$ is less than its mean square deviation from any other function $h(\xi)$:

$$\sqrt{\mathrm{E}[\eta - f(\xi)]^2} \le \sqrt{\mathrm{E}[\eta - h(\xi)]^2}. \qquad (7.9)$$

Proof. Denote $h(\xi) - f(\xi) = u(\xi)$ and note the linearity of expectation, then

$$\mathrm{E}[\eta - h(\xi)]^2 = \mathrm{E}[\eta - f(\xi)] - u(\xi)]^2 =$$

$$\mathrm{E}[\eta - f(\xi)]^2 + \mathrm{E}[u(\xi)]^2 - 2\mathrm{E}[\eta - f(\xi)]u(\xi)$$

but the last term disappears because of (7.6) and

$$\mathrm{E}[\eta - f(\xi)]u(\xi) = \mathrm{E}\eta u(\xi) - \mathrm{E}f(\xi)u(\xi) = 0.$$

Therefore

$$\mathrm{E}[\eta - h(\xi)]^2 = \mathrm{E}[\eta - f(\xi)]^2 + \mathrm{E}[h(\xi) - f(\xi)]^2 \qquad (7.10)$$

hence inequality (7.9).

Property (7.6) only connects functions linear with respect to η; mean values of non-linear functions can change when η is replaced by $f(\xi)$. Thus, there will be an inequality for variances

$$\sigma^2(\eta) = \mathrm{E}[\eta - \mathrm{E}\eta]^2 \ge \sigma^2[f(\xi)]. \qquad (7.11)$$

Indeed, see (7.10),

$$h(\xi) = b = \mathrm{E}\eta = \mathrm{E}f(\xi),$$
$$\mathrm{E}(\eta - b)^2 = \mathrm{E}[\eta - f(\xi)]^2 + \mathrm{E}[f(\xi) - b]^2 \ge \sigma^2[f(\xi)].$$

7.3. Linear correlation

Definition. Correlative dependence between random variables ξ and η is linear if both $f(x)$ and $g(y)$ are linear. In such cases both these functions are called regression (straight) lines.

We derive now the equation of regression line of η on ξ

$$f(x) = Ax + B.$$

Denote $\mathrm{E}\xi = a$, $\mathrm{E}\eta = b$, $\mathrm{E}(\xi - a)^2 = \sigma_1^2$, $\mathrm{E}(\eta - b)^2 = \sigma_2^2$.

First of all by formula (7.3) we determine

$$\mathrm{E}\eta = \mathrm{E}f(\xi) = \mathrm{E}(A\xi + B),\ b = Aa + B,\ B = b - Aa.$$

Then, making use of formula (7.4), we find that

$$\mathrm{E}\xi\eta = \mathrm{E}\xi f(\xi) = \mathrm{E}(A\xi^2 + B\xi) = A\mathrm{E}\xi^2 + (b - Aa)a,$$

$$A = \frac{\mathrm{E}\xi\eta - ab}{\mathrm{E}\xi^2 - a^2} = \frac{\mathrm{E}\xi\eta - ab}{\sigma_1^2} = \rho(\eta/\xi).$$

This coefficient $\rho(\eta/\xi)$ is called the regression coefficient of η on ξ. The regression line of η on ξ is therefore

$$y = \rho(\eta/\xi)(x - a) + b \qquad (7.12a)$$

and similarly

$$x = \rho(\xi/\eta)(y - b) + a, \ \rho(\xi/\eta) = \frac{\mathrm{E}\xi\eta - ab}{\sigma_2^2}. \qquad (7.13a)$$

This ρ is the regression coefficient of ξ on η.

More symmetrically the regression lines can be written by means of a non-dimensional coefficient symmetric with respect to ξ and η

$$r = \frac{\mathrm{E}\xi\eta - ab}{\sigma_1\sigma_2} \qquad (7.14)$$

which is called correlation coefficient between ξ and η. And now we have

$$\rho(\eta/\xi) = r\frac{\sigma_2}{\sigma_1}, \ \ \rho(\xi/\eta) = r\frac{\sigma_1}{\sigma_2}$$

and the regression lines become

$$\frac{y-b}{\sigma_2} = r\frac{x-a}{\sigma_1}, \ \ \frac{x-a}{\sigma_1} = r\frac{y-b}{\sigma_2}. \qquad (7.12b)\ (7.13b)$$

Both these straight lines pass through point (a, b), the centre of the joint distribution of ξ and η. Their slopes are

$$\tan\alpha = r\frac{\sigma_2}{\sigma_1}, \ \ \tan\beta = \frac{\sigma_2}{r\sigma_1}.$$

In § 7.4.1 we prove that $|r| \le 1$ so that $|\tan\alpha| \le |\tan\beta|$. This means that the regression line of η on ξ has a lesser slope than the other regression line. The nearer is $|r|$ to 1, the smaller is the angle between those lines which coincide then and only then when $|r| = 1$.

If $r = 0$, the equations of the regression lines are $y = b$, $x = a$ and $\mathrm{E}_x\eta = b = \mathrm{E}\eta$, $\mathrm{E}_y\xi = a = \mathrm{E}\xi$. The regression coefficients have the same signs as the correlation coefficient r and are connected with it by equation

$\rho(\eta/\xi)\rho(\xi/\eta) = r^2.$ (7.15)

The signs of $\rho(\eta/\xi)$ and $\rho(\xi/\eta)$ coincide which means that, in particular, if η generally increases with ξ, ξ just the same ought to increase generally with η. However, the rapidity of their increase essentially depends on the correlation coefficient.

7.3.1. *Normal correlation.* Correlation between ξ and η is called normal if the density of the two-dimensional distribution of probabilities of (ξ, η) is $(A,\ C > 0, AC - B^2 > 0)$

$$p(x, y) = \frac{\sqrt{Ac - B^2}}{2\pi}\times$$
$$\exp\{-\frac{1}{2}[A(x-a)^2 + 2B(x-a)(y-b) + C(y-b)^2]\}.$$

Then, according to formula (2.35), the particular density of ξ will be

$$\psi_1(x) = \int_{-\infty}^{\infty} p(x, y)dy = \frac{\sqrt{AC - B^2}}{2\pi}\times$$
$$\int_{-\infty}^{\infty} \exp\{-\frac{1}{2}C[(y-b)+\frac{B}{C}(x-a)]^2 - \frac{1}{2}(A - \frac{B^2}{C})(x-a)^2\}dy =$$

$$\frac{\sqrt{AC - B^2}}{\sqrt{2\pi C}}\exp\{-\frac{1}{2}[A - \frac{B^2}{C}](x-a)^2\}$$

since for any λ

$$\sqrt{\frac{C}{2\pi}}\int_{-\infty}^{\infty} \exp[-\frac{1}{2}C(y-\lambda)^2]dy = 1.$$

It is seen now that the particular distribution of ξ is normal with parameters a and

$$\sigma_1^2 = \frac{C}{AC - B^2}.$$

Similarly, the particular distribution of η is normal with parameters b and

$$\sigma_2^2 = \frac{A}{AC - B^2}.$$

The conditional distribution of η for a fixed $\xi = x$ is also normally distributed with density

$$\varphi_x(y) = \frac{p(x, y)}{\psi_1(x)} = \sqrt{\frac{C}{2\pi}} \exp\{-\frac{1}{2}C[(y-b)+\frac{B}{C}(x-a)]^2\}$$

so that its centre is

$E_x\eta = b - (B/C)(x - a)$.

Similarly the centre of the conditional distribution of ξ for a fixed $\eta = y$ is

$E_y\xi = a - (B/A)(y - b)$.

It is seen now that the normal correlation is linear and the regression lines are

$y = -(B/C)(x - a) + b, x = -(B/A)(y - b) + a$

and the regression coefficients are

$\rho(\eta/\xi) = -B/C, \rho(\xi/\eta) = -B/A$.

Formula (7.14) leads therefore to

$$r = -B/\sqrt{AC}. \qquad (7.16)$$

7.4. Correlation coefficient

Let us consider in more detail the correlation coefficient between random variables ξ and η (§ 7.3):

$$r(\xi, \eta) = \frac{E\xi\eta - E\xi E\eta}{\sigma(\xi)\sigma(\eta)}. \qquad (7.14^*)$$

It characterizes the relative deviations of the difference $E\xi\eta - E\xi E\eta$. Such deviations only concern dependent variables and we may therefore say that the correlation coefficient measures the dependence between ξ and η.

Now we can generalize the addition theorem on dependent variables:

$$\sigma^2(\xi + \eta) = \sigma^2(\xi) + \sigma^2(\eta) + 2r(\xi, \eta)\sigma(\xi)\sigma(\eta). \qquad (7.17)$$

It follows from a formula in § 3.3.1 that

$\sigma^2(\xi + \eta) = E(\xi - a)^2 + 2E(\xi - a)(\eta - b) + E(\eta - b)^2$

since

$E(\xi - a)(\eta - b) = E\xi\eta - aE\eta - bE\xi + ab = E\xi\eta - ab = r\sigma_1\sigma_2$.

The correlation coefficient is dimensionless and can therefore be expressed as an expectation of the product of normed and centred deviations

$$\xi_0 = \frac{\xi - a}{\sigma_1},\ \eta_0 = \frac{\eta - b}{\sigma_2}.$$

Indeed,

$$\mathrm{E}\,\xi_0\eta_0 = \mathrm{E}[\frac{\xi - a}{\sigma_1}\frac{\eta - b}{\sigma_2}] = \frac{\mathrm{E}(\xi - a)(\eta - b)}{\sigma_1\sigma_2} = r_0. \qquad (7.18)$$

7.4.1. *The properties of the correlation coefficient*

Theorem 1. Linear transformations of random variables ξ and η do not change the correlation coefficient between them: for any c_1, $c_3 > 0$ and c_2, c_4

$$r(c_1\xi + c_2, c_3\eta + c_4) = r(\xi, \eta).$$

Proof. With $c_1 > 0$ the transformation from ξ to $\xi' = c_1\xi + c_2$ is

$$\mathrm{E}\xi' = c_1\mathrm{E}\xi + c_2 = c_1a + c_2,\ \sigma(\xi') = c_1\sigma(\xi) = c_1\sigma_1,$$

$$\xi'_0 = \frac{\xi' - \mathrm{E}\xi'}{\sigma(\xi')} = \frac{c_1\xi + c_2 - (c_1a + c_2)}{c_1\sigma_1} = \frac{\xi - a}{\sigma_1} = \xi_0.$$

Theorem 2. The domain of the correlation coefficient $r(\xi, \eta)$ is (– 1, 1) and it only takes these extreme values when ξ and η are functionally connected.

Proof. From formula (7.18) and formulas

$$\mathrm{E}\,\xi_0^2 = \frac{\mathrm{E}(\xi - a)^2}{\sigma_1^2} = 1,\ \mathrm{E}\eta_0^2 = 1$$

it follows that

$$\mathrm{E}(\xi_0 \pm \eta_0)^2 = \mathrm{E}\xi_0^2 \pm 2\mathrm{E}\xi_0\eta_0 + \mathrm{E}\eta_0^2 = 1 \pm 2r(\xi, \eta) + 1,$$

$$1 \pm r(\xi, \eta) = \frac{1}{2}\mathrm{E}(\xi_0 \pm \eta_0)^2 \geq 0, \qquad (7.19)$$

$$-1 \leq r(\xi, \eta) \leq 1.$$

Equality in (7.19) takes place when and only when

$\mathrm{E}(\xi_0 \pm \eta_0)^2 = 0$; that is when $\xi_0 \pm \eta_0 = 0$ which means that

$$\frac{\xi - a}{\sigma_1} \pm \frac{\eta - b}{\sigma_2} = 0,\ \ \eta = b\,\mathrm{m}\frac{\sigma_2}{\sigma_1}(\xi - a).$$

Theorem 3. The correlation coefficient between independent variables disappears. This directly follows from (3.9) and (7.14*).

Note that the inverse proposition is not valid but that when $r(\xi, \eta) = 0$, random variables ξ and η are called uncorrelated. In one important case they are independent; it occurs when the correlation is normal.

Indeed, as shown by formula (7.16), the coefficient $r(\xi, \eta) = 0$ then and only then when $B = 0$. But then

$$p(x, y) = \frac{\sqrt{AC}}{2\pi}\exp\{-\frac{1}{2}[A(x-a)^2 + C(y-b)^2]\} =$$

$$\frac{\sqrt{A}}{\sqrt{2\pi}}\exp[-\frac{1}{2}A(x-a)^2]\frac{C}{\sqrt{2\pi}}\exp[-\frac{1}{2}C(y-b)^2],\ \text{QED.}$$

7.5. The best linear approximation to the regression function

For linear correlation which we were discussing the parameters of the regression function are determined comparatively easy. In cases of more complicated correlative dependence the derivation of that function is considerably difficult, hence the problem of its best approximation. In § 7.2 we have established that, for any $h(\xi)$, $f(\xi)$ ensures the best mean square approximation to η:

$$\mathrm{E}[\eta - f(\xi)]^2 \le \mathrm{E}[\eta - h(\xi)]^2.$$

It is therefore natural to determine the linear function $(Ax + B)$ best approximating the regression function $f(x)$ as such for which

$$\mathrm{E}[\eta - (A\xi + B)]^2 = \min.$$

It occurs that the parameters A and B can be determined just like the parameters of the linear regression function were (§ 7.3). More specifically:

Theorem. The mean square deviation of random variable η from $(A\xi + B)$ is minimal then and only then when

$$A = \rho(\eta/\xi) = r\sigma_2/\sigma_1,\ B = b - Aa,\ Ax + B = r(\sigma_2/\sigma_1)(\xi - a) + b.$$

Consequently, among all the straight lines and for any correlative dependence the regression line (7.12a) ensures the best approximation in the mean to the actual regression of η on ξ.

Proof. Denote $B - (b - Aa) = C$, then, taking into account equalities $\mathrm{E}(\xi - a) = \mathrm{E}(\eta - b) = 0$, transform

$$\mathrm{E}[\eta - (Ax + B)]^2 = \mathrm{E}[(\eta - b) - A(\xi - a) - C]^2 =$$

$$\mathrm{E}(\eta - b)^2 + A^2\mathrm{E}(\xi - a)^2 - 2A\mathrm{E}(\xi - a)(\eta - b) + C^2 =$$

$$\sigma_2{}^2 + A^2\sigma_1{}^2 - 2Ar\sigma_1\sigma_2 + C^2. \qquad (7.20)$$

Here, $\sigma_2{}^2$ is constant, $C^2 = 0$ if $B = b - Aa$ and

$$A^2\sigma_1{}^2 - 2Ar\sigma_1\sigma_2 = \sigma_1{}^2(A - r\sigma_2/\sigma_1)^2 - r^2\sigma_2{}^2$$

takes its minimal value $- r^2\sigma_2{}^2$ when $A = r\sigma_2/\sigma_1$, QED.

We will additionally determine the mean square deviation of η from $r(\sigma_2/\sigma_1)(\xi - a) + b$ which ensures the best linear approximation in the mean. From formula (7.20) taking $C = 0$ and $A = r\sigma_2/\sigma_1$, we have

$$\mathrm{E}\{\eta - [r(\sigma_2/\sigma_1)(\xi - a) + b]\}^2 = \sigma_2{}^2 - r^2\sigma_2{}^2 = \sigma_2{}^2(1 - r^2). \quad (7.21)$$

This formula, since it determines $\sqrt{1-r^2}$, describes how the correlation coefficient characterizes dependence And so, $r(\xi, \eta)$ characterizes the relative magnitude of the mean square deviation in the left side of (7.21) and therefore the measure of the linear connection between ξ and η. The nearer is r^2 to 1, the less in the mean is the scatter of the values of η relative to the regression line of η on ξ. And all the above is certainly valid with respect to the regression of ξ on η.

7.6. Analysing linear correlation by random sampling. The significance of the correlation coefficient

For analysing linear correlation between ξ and η independent trials are made; the outcome of each is a pair of numbers (x_i, y_i). When considering n such pairs as a random sample from the population of all possible values of (ξ, η), we can find approximate values of all the parameters of the linear correlation between ξ and η by the method of moments (Chapter 4). First of all, we have the following approximate formulas

$$a = \mathrm{E}\xi \approx \bar{x} = \frac{\sum x}{n}; \; b = \mathrm{E}\eta \approx \bar{y} = \frac{\sum y}{n};$$

$$\sigma^2(\xi) \approx s_1^2 = \frac{\sum (x-\bar{x})^2}{n-1}; \; \sigma^2(\eta) \approx s_2^2 = \frac{\sum (y-\bar{y})^2}{n-1}; \qquad (7.22)$$

$$\mathrm{E}(\xi - a)(\eta - b) \approx \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1}. \qquad (7.23)$$

We can now derive an approximate formula for the correlation coefficient

$$r(\xi, \eta) \approx r_n = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)s_1 s_2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} \quad (7.24)$$

where r_n is the sample correlation coefficient.

Replacing in (7.12b) and (7.13b) all the expectations by the corresponding mean values, we get the sample regression lines of η on ξ and ξ on η:

$$y-\overline{y}=r_n\frac{s_2}{s_1}(x-\overline{x}),\ x-\overline{x}=r_n\frac{s_1}{s_2}(y-\overline{y}).\qquad (7.25;\ 7.26)$$

The coefficients r_ns_2/s_1 and r_ns_1/s_2 are called sample regression coefficients. It is important to note that the sample magnitudes (7.25) and (7.26) have the property similar to that discussed in § 7.5; the sum of the squares of the deviations of the observed values y_i from the sample regression line (7.25) is less than from any other straight line:

$$\sum_{i=1}^{n}\{y_i-[\overline{y}+r_n\frac{s_2}{s_1}(x-\overline{x})]\}^2\le\sum_{i=1}^{n}[y_i-(Ax_i+B)]^2.$$

The proof is similar to that carried out in § 7.5. The same can be stated about the regression line (7.26).

Formulas (7.22) – (7.26) show that the determination of the sample regression lines requires approximate calculations with a large number of the differences $(x_i-\overline{x})$ and $(y_i-\overline{y})$. Just like in § 6.2.1 these calculations can be essentially simplified by a preliminary linear transformation of x and y:

$u = (x - x_0)/h_1$, $v = (y - y_0)/h_2$, $h_1, h_2 > 0$.

We will have […]

Example. […]

7.6.1. *A note on the confidence probability of the correlation coefficient*. This problem is beyond our scope. We only note that the three-sigma rule is not recommended here since even with a large n the distribution of probabilities of the sample correlation coefficient considerably differs from normality.

We restrict our attention to a simpler problem: can it happen that the sampling correlation coefficient accidentally differs from zero whereas the random variables ξ and η are not really correlated? The solution of this problem assumes that the real correlation coefficient $r(\xi, \eta) = 0$. We provide a table of the boundaries of random deviations of the product $|r|\sqrt{n-1}$ from zero depending on the assigned probability P and n as well as on an additional condition that the studied correlation little differs from normality.

If for the sample correlation coefficient r_n the product mentioned above exceeds the boundary value given in the table, we may state with probability P that the real correlation coefficient $r(\xi, \eta)$ differs from zero. […]

Example. […]

7.7. Exercises

1) Calculate the correlation coefficient between λ_1 and λ_2 in Exercise 1 to Chapter 3. *Answer*:

$$r(\lambda_1, \lambda_2) = \frac{E\lambda_1\lambda_2 - E\lambda_1 E\lambda_2}{\sigma_1\sigma_2} = [...]$$

2) Determine the linear correlation by issuing from sample observations […]

Notes

1. Concerning true values see Sheynin (2007).

2. Complete group of events: Events A_1, A_2, …, A_n comprise a complete group if they are pairwise incompatible and event ($A_1 + A_2 + … + A_n$) is certain. This notion is widely used when deriving many stochastic propositions (Vatutin 1999). Rumshitsky, however, seems to make too much of this notion.

3. This is Chebyshev's definition (1845/1951, p. 29) of the aim of the probability theory.

4. This example is not quite proper: $P(A$ and $B) = 8.25/100$.

5. This example is due to Bernstein (1946, p. 47). Feller (1950/1964, § 6.3) offered another example.

6. Centre of a random variable, of a distribution was introduced in 1970 on a much more sophisticated level (Kruglov 1999). Again (see Note 2), Rumshitsky seems to make too much of this notion.

7. In other words, study a sample.

8. Jakob Bernoulli also investigated the rapidity of the convergence of frequency to probability. Below, the author does not mention the Poisson law of large numbers.

9. Sampling deserved more attention.

10. Laplace had indeed developed De Moivre's result in Chapter 3 of his *Théorie analytique des probabilités* (and Markov coined the term *De Moivre – Laplace theorem*), but the author's reference to 1783 is wrong.

11. This is wrong. Chebyshev had not proved the CLT rigorously (Gnedenko & Sheynin 1978/2001, p. 260).

12. It is worthwhile to quote Lindeberg (1922, p. 211):

Nunmehr finde ich, daß schon Liapunov […] *allgemeine Resultate dargelegt hat, die nicht nur über diejenigen des Herrn v. Mises hinausgehen, sondern aus denen auch die meisten der von mir bewiesen Tatsachen abgeleitet werden können.*

13. This chapter is unworthy. Direct and indirect measurements denote the determination of one or several unknowns respectively. The author offered his own (unfortunate) definition of direct measurements and did not mention the other case (therefore, omitted the method of least squares). He superficially described systematic errors and (below) wrongly stated that random errors invariably obey the normal distribution.

14. The three-sigma rule presumes normally distributed errors which the author only mentioned above (cf. Note 13). See also his pertinent remark in § 7.6.1.

15. Here and below the author retained too many significant digits.

Bibliography

Arley N., Buch K. R. (1949, 1950), *Introduction to the Theory of Probability and Statistics*. New York.

Bernstein S. N. (1946), *Teoria Veroiatnostei* (Theory of Probability). Moscow – Leningrad. 4th edition.

Bertrand J. (1888), *Calcul des probabilités*. Paris, 1907. Bronx, N. Y., 1970, 1972.

Chebyshev P. L. (1845), *Opyt Elementarnogo Analysa Teorii Veroiatnostei* (Essay on Elementary Analysis of Probability Theory). *Polnoe Sobr. Soch.* (Complete Works), vol. 5. Moscow – Leningrad, 1951, pp. 26 – 87.

Cramér H. (1946), *Mathematical Methods of Statistics.* Princeton, 1999.

Feller W. (1950), *Introduction to Probability Theory and Its Applications*, vol. 1. New York – London, 1957, 1968. Russian translation: Moscow, 1964.

Gnedenko B. V. (1950), *Kurs Teorii Veroiatnostei*. Moscow, 1954, 1969, 2001. Translations: *Theory of Probability*. Moscow, 1969, 1973. *Lehrbuch der Wahrscheinlichkeitsrechnung*. Many editions: Berlin, 1957 – 1987.

Gnedenko B. V., Khinchin A. Ya. (1946, in Russian), *Elementary Introduction to the Theory of Probability*. Numerous editions up to 2012. English translation: Mineola, N. Y., 1962.

Gnedenko B. V., Sheynin O. (1978, in Russian), Theory of probability. In *Mathematics of the 19th Century*, vol. 1. Editors, A. N. Kolmogorov, A. P. Youshkevich. Basel, 1992, 2001, pp. 211 – 288.

Khinchin A. Ya. (1927), Über das Gesetz der großen Zahlen. *Math. Ann.*, Bd. 96, pp. 152 – 168.

Kruglov V.M. (1999), Centre of a random variable. In Prokhorov (1999, p. 794).

Lindeberg J. W. (1922), Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.*, Bd. 15, pp. 211 – 225.

Prokhorov Yu. V., Editor (1999), *Veroiatnost i Matematicheskaia Statistika. Enziklopedia* (Probability and Mathematical Statistics. Encyclopedia). Moscow.

Romanovsly V. I. (1947), *Osnovnye Zadachi Teorii Oshibok* (Main Problems of the Theory of Errors). Moscow – Leningrad.

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

Vatutin V. A. (1999), Disjoint events. In Prokhorov (1999, p. 400).