THE TEACHING OF MATHEMATICS
1998, Vol. I, pp. 1–22

SELECTED CHAPTERS FROM ALGEBRA

I. R. Shafarevich

Abstract. The aim of this publication (this paper together with several its continuations) is to present algebra as a branch of mathematics treating the contents close to the usual teaching matter. The whole exposition presupposes not a large frame of knowledge: operations with integers and fractions, square roots, removing of parentheses and other transformations of literal terms, properties of inequalities. The exposition clusters round a number of main themes: "Number", "Polynomial", "Set", each of which is treated in a series of chapters listed in Preface.

Preface

In the school mathematical education algebra has the role of Cinderella and geometry of Beloved Daughter. The extent of the geometrical knowledge, studied in school, coincides approximately with the development in this field attained in Ancient Greece and embodied in Euclid's "Elements" (III century B.C.). For long, geometry had been taught after Euclid and, only at a later time, some simplified versions appeared. In spite of all changes introduced into geometry course, the influence of Euclid and the spirit of the grandiose scientific revolution of Hellenic era continued to last. More than once I met people saying: "I have not chosen mathematics to be my profession, but I will remember forever all the beauty of logical construction of geometry with the precise derivations of more and more complex statements starting with the simplest".

Unfortunately, not even once I heard a similar reaction concerning algebra. The school course of algebra is a strange mixture of useful rules, logical reasonings, practices of how to use such auxilliary tools as tables of logarithms or a microcalculator. In its spirit, such a course is closer to the type of mathematical knowledge formed in Ancient Egypt or Babylon than to the direction of development which started in Ancient Greece and was continued in Western Europe, in the Renaissance period. None the less, algebra is a fundamental, deep and beautiful branch of mathematics as much as geometry is. Moreover, from the point of view of the contemporary classification of mathematics, the school course of algebra contains the elements of several subdivisions of mathematics: algebra, number theory, combinatorics and a small part of probability theory.

This paper is an English translation of: И. Р. Шафаревич, *Избраные главы алгебры*, Математическое образование, 1, 1, апр.–јун 1997, Москва, стр. 5–27. In the opinion of the editors, the paper merits wider circulation and we are thankful to the author for his kind permission to let us make this version.

The aim of this publication (this paper together with several its continuations) is to exhibit algebra as a branch of mathematics treating the contents close to the usual teaching matter. The whole exposition resupposes not a large frame of knowledge: operations with integers and fractions, square roots, cancellation of brackets and other transformations of literal terms, properties of inequalities. And all these practices are very well settled until the 9th class. The complexity of mathematical reasonings somewhat increases as we proceed with the matter. In order to help the reader digest the text with more ease, we also include some simple exercises.

The exposition clusters round a number of the main themes: "Number", "Polynomial", "Set", each of which is treated in more than one chapter, and the chapters related to different themes do not overlap. In the form of appendices, some more complex questions are selected, which are related to the rest of the text and which comprise no new facts besides those already in the reader's mind. In the first chapters they do not appear.

An expected list of chapters:

Chapter 1. Number.

(Irrationality of $\sqrt{2}$ and other radicals. Unique factorization of a positive integer as the product of primes.)

Chapter 2. Polynomial.

(Roots and linear factors. Common roots. Interpolation. Multiple roots. Derivative of a polynomial. Newton's binomial.)

Chapter 3. Set.

(Finite sets and their subsets. Combinatorics. Some concepts from probability theory.)

Chapter 4. Number (continued).

(Axioms of real numbers. Properties of polynomials as continuous functions.)

Chapter 5. Polynomial (continued).

(Separation of roots of a polynomial. Sturm's theorem.)

Chapter 6. Set (continued).

(Infinite sets, countable and uncountable sets.)

Chapter 7. Number (continued).

(Infinite set of prime numbers. Density of the set of prime numbers.)

Appendix I.

(Chebyshev's estimations of the number of primes less than the given bound.)

Chapter 8. Number (continued).

(Complex numbers.)

Chapter 9. Polynomial (continued).

(The existence of complex root of a polynomial with complex coefficients.)

Chapter 10. Number (continued).

(Arithmetic of Gauss numbers, number-theoretical applications.)

Chapter 11. Polynomial (continued).

(Constructions by means of compasses and ruler and the solution of equations by square radicals.)

Chapter 12. Polynomial (continued).

(Symmetric functions.)

Chapter 13. Polynomial (continued).

(Solutions of cubic and biquadratic equations. Nonsolvability by radicals of equations of degree $n \geqslant 5$.)

Appendix II.

(Equations of degree 5, icosahedron, problem of resolvents.)

Chapter 14. Number (ended).

(Finite fields and finite geometries.)

Appendix III.

(Construction of regular 17-gon.)

Chapter 15. Polynomial (ended).

(Formal power series and infinite products. Applications to number theory.)

CHAPTER I. NUMBER

1. Irrational numbers

Natural numbers arose as a result of counting. The cognition of the fact that two eyes, two men walking side by side and two oars of a boat have something in common, expressed by the abstract concept “two”, was an important step in the logical development of mankind. The step to follow was not made so easily. Consonance, in many languages, of the word “three” and the words “many” (orig. “много”) or “much” (orig. “слишком”) bears a record to it. And only step by step, the idea of infinite series of natural numbers arose.

Gradually, the concept of number was related not only to counting, but also to measuring of length, area, weight etc. To be more concrete, we will only consider the length of line segments in the following. First of all, we have to choose a unit of length: cm, mm, km, light-year, ... Thus, a segment E is fixed and may be used for measuring of another segment A. If E is contained in A exactly n times, then we say that the length of segment A is equal to n (Fig. 1,a). But, as a rule, it will not happen (Fig. 1,b).

a) b)

Fig. 1

Then, the unit lessens, breaking up E into m equal segments E'. If E' is contained in A exactly n times, then we say that the length of A is equal to $\frac{n}{m}$ (relative to the unit E). Thousands of years, men, in different parts of the world, had applied this procedure in a variety of situations until the question: *is this breaking really possible?*, arose. This completely new setting of question already belongs to the historical epoch—Pythagoras' School, in the period of VI or V century B.C. The segments A and E are called *commensurable* if there exists a segment E' exactly m times contained in E and n times in A. Thus the above question modifies to the following: *are each two segments commensurable?* Or further: *is the length of each segment* (a unit being fixed) *equal to a rational number* $\frac{n}{m}$? The answer is *negative* and an example of a pair of incommensurable segments is simple. Consider a square, its side being E and its diagonal A.

Theorem 1. *The side of the square is incommensurable with its diagonal.*

Before we proceed with the proof, we give this theorem another form. According to the famous Pythagorean theorem, the area of the square over the hypotenuse of a right triangle is equal to the sum of areas of the squares over the other two sides. Or, in other words, the square of the length of the hypotenuse is equal to the sum of squares of the lengths of the other two sides. However, the diagonal A of our square is the hypotenuse of the isosceles triangle whose other two sides coincide with the sides E of the square (Fig. 2) and hence in our case $A^2 = 2E^2$, and if $A = nE'$, $E = mE'$, then $\left(\frac{n}{m}\right)^2 = 2$ or $\frac{n}{m} = \sqrt{2}$. Therefore, Theorem 1 can be reformulated as

Theorem 2. *$\sqrt{2}$ is not a rational number.*

Fig. 2 Fig. 3

We shall give a proof of this form of the theorem, but first we make the following remark. Although we have leaned on the Pythagorean theorem, we have actually used it only for the case of an isosceles right triangle, when the conclusion is evident. Namely, it is enough to complete the Fig. 2 by constructing the square over the side A (Fig. 3). From known criterions of congruency it follows that all five small right isosceles triangles in Fig. 3 are congruent. Hence they have the same area S. But the square whose side is E consists of two such triangles and its area is E^2. Thus, $E^2 = 2S$. Similarly, $A^2 = 4S$. Hence, $A^2 = 2E^2$, i.e. $(A/E)^2 = 2$, which is all we need.

We can now proceed with the proof of Theorem 2. Since our task is to prove the *impossibility* of representing $\sqrt{2}$ in the form $\sqrt{2} = \frac{n}{m}$, it is natural to start with the converse, i.e. to suppose that $\sqrt{2} = \frac{n}{m}$, where n and m are positive integers. We also suppose that they are relatively prime, for if they have a common factor, it can be cancelled without changing the ratio $\frac{n}{m}$. By definition of the square root, the equality $\sqrt{2} = \frac{n}{m}$ means that $2 = \left(\frac{n}{m}\right)^2 = \frac{n^2}{m^2}$. Multiplying both sides by m^2 we obtain the equality

$$2m^2 = n^2, \tag{1}$$

where m and n are relatively prime positive integers, and it remains to prove that it is impossible.

Since there is a factor 2 on the left-hand side of (1), the question is naturally related to the possibility of dividing positive integers by 2. Numbers divisible by 2 are said to be *even*, and those indivisible by 2 to be *odd*. Therefore, every even number k can be written in the form $k = 2l$, where l is a positive integer, i.e. we have an explicit expression for even numbers, whereas odd numbers are defined only by a negative statement—that such an expression does not hold for them. But it is easy to obtain an explicit expression for odd numbers.

Lemma 1. *Every odd number r can be written in the form $r = 2s + 1$, where s is a natural number or 0. Conversely, all such numbers are odd.*

The last statement is evident: if $r = 2s + 1$ were even, it would be of the form $r = 2l$, which implies $2l = 2s + 1$, i.e. $2(l - s) = 1$, which is a contradiction.

In order to prove the first statement, notice that if the odd number $r \leqslant 2$, then $r = 1$ and the representation is true with $s = 0$. If the odd number $r > 1$, then $r \geqslant 3$. Subtracting 2 from it, we obtain the number $r_1 = r - 2 \geqslant 1$, and r_1 is again odd. If it is greater then 1, we again subtract 2 and put $r_2 = r_1 - 2$. In this way we obtain a decreasing sequence of numbers r, r_1, r_2, ... , where each member is less than its predecessor by 2. We continue this procedure as long as $r_i \geqslant 1$, and since positive integers cannot decrease indefinitely, we shall arrive at the situation when we cannot further subtract the number 2, i.e. when $r_i = 1$. We obtain that $r_i = r_{i-1} - 2 = r_{i-2} - 2 - 2 = \cdots = r - 2 - 2 \cdots - 2 = r - 2i = 1$. Hence $r = 2i + 1$, as stated.

We can now prove the basic property of even and odd numbers.

Lemma 2. *The product of two even numbers is even, the product of an even and an odd number is even and the product of two odd numbers is odd.*

The first two statements follow directly from the definition of even numbers: if $k = 2l$, then no matter whether m is even or odd, we always have $km = 2lm$ which is an even number. However, the proof of the last statement requires Lemma 1. Let k_1 and k_2 be two odd numbers. By Lemma 1 we can write them in the form $k_1 = 2s_1 + 1$, $k_2 = 2s_2 + 1$, where s_1 and s_2 are natural numbers or 0. Then $k_1k_2 = (2s_1+1)(2s_2+1) = 4s_1s_2+2s_1+2s_2+1 = 2s+1$ where $s = 2s_1s_2+s_1+s_2$. As we know, any number of the form $2s + 1$ is odd and so k_1k_2 is odd.

Notice the following particular case of Lemma 2: the square of an odd number is odd.

Now we can easily finish the proof of Theorem 2. Suppose that the equality (1) is true where m and n are positive integers, relatively prime. If n is odd, then by Lemma 2, n^2 is also odd, whereas from (1) follows that n^2 is even. Hence, n is even and can be written in the form $n = 2s$. But m and n are relatively prime, and so m must be odd (otherwise they would have common factor 2). Substituting the expression for n into (1) and cancelling by 2, we obtain

$$m^2 = 2s^2,$$

i.e. the square of the odd number m is even, which is in contradiction with Lemma 2. Theorem 2, and hence Theorem 1, are proved.

Under the supposition that the result of measuring the length of a segment (with respect to a given unit segment) is a number and that the square root of a positive number is a number, we can look at Theorems 1 and 2 from a different point of view. These theorems assert that in the case of the diagonal of a square or in the case of $\sqrt{2}$, these numbers are not rational, that is to say they are irrational. This is the simplest example of an irrational number. All the numbers, rational and irrational, comprise real numbers. In one of the following chapters we shall give a more precise logical approach to the concept of a real number, and we shall use it in accordance with the school teaching of mathematics, that is to say we shall not insist too much on the logical foundations.

Why did such a simple and at the same time important fact, the existence of irrational numbers, have to wait so long to be discovered? The answer is simple—because for all practical purposes, we can take, for instance, $\sqrt{2}$ to be a rational number. In fact, we have

Theorem 3. *No matter how small is a given number ε, it is possible to find a rational number $a = \dfrac{m}{n}$, such that $a < \sqrt{2}$ and $\sqrt{2} - a < \varepsilon$.*

All practical measurements can necessarily be carried out only up to a certain degree of accuracy, and up to that degree of accuracy we may take $\sqrt{2}$ to be rational. Hence, we can say that our measurements give us $\sqrt{2}$ as a rational number.

In order to prove Theorem 3 it is enough to write our arbitrarily small number ε in the form $\dfrac{1}{10^n}$ for sufficiently large n, and to find a positive integer k such that

$$\frac{k}{10^n} \leqslant \sqrt{2} < \frac{k+1}{10^n}. \tag{2}$$

Then we may take $a = \dfrac{k}{10^n}$, for $\sqrt{2} - \dfrac{k}{10^n} < \dfrac{1}{10^n}$. The inequalities (2) are equivalent to $\dfrac{k^2}{10^{2n}} \leqslant 2 < \dfrac{(k+1)^2}{10^{2n}}$ or $k^2 \leqslant 2 \cdot 10^{2n} < (k+1)^2$. Since the number n, and so $2 \cdot 10^n$, is given, there exists the largest positive integer k whose square is not greater than $2 \cdot 10^n$. This is the number we want.

Obviously, the conclusion of Theorem 3 holds not only for the number $\sqrt{2}$, but also for any positive (for simplicity's sake we confine ourselves to them) real number x. This becomes evident if we represent x by a point of the number axis, if we divide the unit length E into small segments $\dfrac{1}{10^n} E$ and cover the entire line by these segments (Fig. 4).

Fig. 4

Then the last of those points which is not right from x gives the required rational number: if it is the k-th point, then $a = \dfrac{k}{10^n} \leqslant x$ and $x - a < \dfrac{1}{10^n}$.

Now please consider the depth of the assertion contained in Theorems 1 and 2. This assertion can *never* be verified by an experiment, since an experiment can be carried out up to a certain degree of accuracy, and $\sqrt{2}$ can be expressed as a rational number with any given degree of accuracy. It is an accomplishment of *pure reasoning* which could not be achieved even as a result of thousands of years of experience. It had to wait for the revolution in mathematics carried out in Ancient Greece in VII–V centuries B.C. No wonder that in the Pythagorean School those facts were considered to be holy, secret knowledge, not to be shared with ordinary people. The legend says that Hyppas, a Pythagorean, died in a shipwreck as a punisment for revealing this secret. A hundred years later Plato in his book "Laws" narrates how he was astonished when he found that it is not always possible "to measure a length by a length". He speaks of his "shameful ignorance": "It seemed to me that it is not appropriate for men, but rather for swine. And I was ashamed not only for myself, but for all Greeks."

The Theorems 1 and 2 may throw some light onto the question often posed to mathematics: why prove theorems? The answer that first comes to mind is: in order to be sure that a statement is true. But sometimes a statement has been verified in so many particular cases that no one doubts its truth (and physicists often snigger at mathematicians who prove undoubted truths). But we have seen that a proof sometimes leads mathematicians into a completely new world of mathematical ideas and concepts, which would not be discovered otherwise.

Problems

1. Prove that the numbers $\sqrt{6}$ and $\sqrt[3]{2}$ are irrational.

2. Prove that the number $\sqrt{2} + \sqrt{3}$ is irrational.

3. Prove that the number $\sqrt[3]{3} + \sqrt{2}$ is irrational.

4. Determine $\sqrt{2}$ with accuracy not less than $\frac{1}{100}$.

5. Prove that every positive integer can be written as a sum of terms of the form 2^k, where all the terms are different. Prove that for each number this representation is unique.

2. Irrationality of other square roots

It would be interesting to generalize the results of the preceding section. For example, is it possible to prove in the same way that $\sqrt{3}$ is irrational? Clearly, we have to adapt the reasoning from the previous section to the new situation.

We want to prove the impossibility of the equality $3 = \left(\frac{n}{m}\right)^2$, or

$$3m^2 = n^2, \tag{3}$$

where, as in section **1**, we may take that the fraction $\frac{n}{m}$ cannot be further cancelled, i.e. that the positive integers m, n are relatively prime. Since in (3) we have the number 3, it is natural to examine the properties of division by 3. We adapt Lemmas 1 and 2 to the new case.

Lemma 3. *Every positive integer r is either divisible by 3 or it can be represented in one of the following forms: $r = 3s + 1$ or $r = 3s + 2$, where s is a natural numbers or 0. The numbers $3s + 1$ and $3s + 2$ are not divisible by 3.*

The last statement is evident. If, for example, $n = 3s + 1$ were divisible by 3, we would have $3s + 1 = 3m$, i.e. $3(m - s) = 1$, which is a contradiction. If $n = 3s + 2$ were divisible by 3 we would have $3s + 2 = m$. i.e. $3(m - s) = 2$, again a contradiction. We prove the first statement of Lemma 3 by the same procedure used to prove Lemma 1. If r is not divisible by 3 and is less than 3, then $r = 1$ or $r = 2$ and the given representation holds with $s = 0$. If $r > 3$, then subtracting 3 from it we get $r_1 = r - 3 > 0$ and r_1 is again not divisible by 3. We continue to subtract the number 3 and we obtain the sequence r, $r_1 = r - 3$, $r_2 = r - 3 - 3$, $\dots$, $r_s = r - 3 - 3 - \cdots - 3$, where we cannot subtract 3 any more, since as noticed above, $r_s = 1$ or $r_s = 2$. As a result we have two possibilities: $r - 3s = 1$, i.e. $r = 3s + 1$ or $r - 3s = 2$, i.e. $r = 3s + 2$, as stated.

In the formulation of the following lemma we take from the formulation of Lemma 2 only that part which we shall use later.

Lemma 4. *The product of two positive integers not divisible by 3 is itself not divisible by 3.*

Let r_1 and r_2 be two positive integers not divisible by 3. According to Lemma 3 for each one there are two possibilities: the number can be written in the form $3s+1$ or in the form $3s + 2$. Hence, there are altogether four possibilities:

1) $r_1 = 3s_1 + 1,\ r_2 = 3s_2 + 1;$ 2) $r_1 = 3s_1 + 1,\ r_2 = 3s_2 + 2;$

3) $r_1 = 3s_1 + 2,\ r_2 = 3s_2 + 1;$ 4) $r_1 = 3s_1 + 2,\ r_2 = 3s_2 + 2;$

The cases 2) and 3) differ only in the numerations r_1 and r_2 and it is enough to consider only one of them (for instance, 2). For the remaining three cases we multiply out:

$$1)\ r_1r_2 = 9s_1s_2 + 3s_1 + 3s_2 + 1 = 3t_1 + 1,\ \ t_1 = 3s_1s_2 + s_1 + s_2;$$

$$2)\ r_1r_2 = 9s_1s_2 + 6s_1 + 3s_2 + 2 = 3t_2 + 2,\ \ t_2 = 3s_1s_2 + 2s_1 + s_2;$$

$$3)\ r_1r_2 = 9s_1s_2 + 6s_1 + 6s_2 + 4 = 3t_3 + 1,\ \ t_3 = 3s_1s_2 + 2s_1 + 2s_2 + 1$$

(in the last formula we put $4 = 3 + 1$, and group 3 with the numbers divisible by 3). As a result we obtain numbers of the form $3t + 1$ and $3t + 2$ which are not divisible by 3 (Lemma 3).

Now we can easily carry over Theorem 2 to our case.

Theorem 3. *$\sqrt{3}$ is not a rational number.*

The proof follows closely the lines of the proof of Theorem 2. We have to establish a contradiction starting with the equality (3): $3m^2 = n^2$, where m and n are relatively prime. If the number n is not divisible by 3, according to Lemma 4 its square is also not divisible by 3. But it is equal to $3m^2$, which means that n is divisible by 3: $n = 3s$. Substituting this into (3) and cancelling by 3 we get $m^2 = 3s^2$. But since n and m are relatively prime and n is divisible by 3, m cannot be divisible by 3. In view of Lemma 4 its square is also not divisible by 3, but it is equal to $3s^2$. This contradiction proves the theorem.

The close parallel between the reasonings used in the two cases we proved leads us to think that we can carry on. Of course, we do not consider $\sqrt{4}$, since $\sqrt{4} = 2$, but we may apply the same line of reasoning to $\sqrt{5}$. Clearly, we shall have to prove a lemma analogous to Lemmas 2 and 4, but the number of products which have to be evaluated will increase. We can verify all of them and conclude that $\sqrt{5}$ is irrational. We can continue and consider $\sqrt{6}$, $\sqrt{7}$, etc. In each new case the number of checkings in the proof of the lemma which corresponds to Lemmas 2 and 4 will increase. Considering all natural numbers n, for instance up to 20, we can conclude that $\sqrt{n}$ is irrational, except in those cases when n is the square of an integer ($n = 4$, 9 and 16). In this way, having to do more and more calculations, we can infer that $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, $\sqrt{6}$, $\sqrt{7}$, $\sqrt{8}$, $\sqrt{10}$, $\sqrt{11}$, $\sqrt{12}$, $\sqrt{13}$, $\sqrt{14}$, $\sqrt{15}$, $\sqrt{17}$, $\sqrt{18}$ and $\sqrt{19}$ are irrational numbers. This leads to the following conjecture: $\sqrt{n}$ is irrational for all positive integers n which are not squares of positive integers. But we cannot prove this general conjecture by the reasoning applied up to now, since in one step of the proof we have to analyse all possible cases.

It is interesting that the road covered by our reasoning was actually covered by mankind. As we have said, the irrationality of $\sqrt{2}$ was proved by the Pythagoreans. Later on irrationality of $\sqrt{n}$ was proved for some relatively small numbers n, until the general problem was formulated. About its solution we read in Plato's dialogue "Theaetetus", written in about 400 B.C. The author narrates how the famous philosopher Socrates met with Theodor, mathematician from Cyrene and his young, very talented pupil by the name of Theaetetus. Theaetetus had the age of today's schoolboy, between 14 and 15 years. Theodor's comment on his abilities reads: "he approaches studying and research with such ease, fluency, eagerness and peace as oil flows form a pot and I wonder how can one achieve so much at that age". Further on, Theaetetus himself informs Socrates about the work he did with a friend, also called Socrates, a namesake of the philosopher. He says that Theodor informed them about the incommensurability (to use contemporary terms) of the side of a square with the unit segment if the area of the square is an integer, but not the square of an integer. If this area is n, this means that $\sqrt{n}$ is irrational. Theodor proved this for $n = 2, 3, 5$, and "solving one case after the other, he came up to 17". Theaetetus became intersted in the problem and, together with his friend Socrates, solved it, as it is recorded at the end of the dialogue. We shall not go into the reasoning of Theodor (there are several hypotheses), but we shall give the proof of the general statement, following the exposition of Euclid which is, very probably analogous to the proof of Theaetetus (with a simplification given by Gauss 2000 years later).

We first prove an analog of Lemmas 1 and 3.

Theorem 4. *For any two positive integers n and m there exist integers t and r, positive or 0, such that $r < m$ and*

$$n = mt + r. \tag{4}$$

For given n and m this representation is unique.

Representation (4) is called *division with remainder of n by m*, the number t is the *quotient* and r is the *remainder.*

The proof follows the known line of reasoning. If $m > n$, the equality (4) holds with $t = 0$, $r = n$. If $n \geqslant m$, then put $n_1 = n - m$. Clearly, $n_1 \geqslant 0$. If $n_1 \geqslant m$, put $n_2 = n_1 - m$. We keep on subtracting m until we arrive at the number $n_t = n - m - \cdots - m = r$, where $r \geqslant 0$ but $< m$. Thus we obtain the required representation $n - mt = r$, i.e. $n = mt + r$.

We now prove its uniqueness for given n and m. Let

$$n = mt_1 + r_1, \quad n = mt_2 + r_2.$$

Let $t_1 \neq t_2$, e.g. $t_1 > t_2$. Subtracting the second equality from the first we get: $m(t_1 - t_2) + r_1 - r_2 = 0$, i.e. $m(t_1 - t_2) = r_2 - r_1$. Since $r_1 < r_2$ on the right-hand side we have a positive number which is less than m, and on the left-hand side a number divisible by m. This is impossible.

Before we prove an analog of Lemmas 2 and 4 we have to introduce (or rather to recall) an important concept.

A positive integer, different from 1, is said to be *prime* if it is divisible only by itself and by 1. For example, among the first twenty numbers the following are prime: 2, 3, 5, 7, 11, 13, 17, 19.

Although obvious, the following property is important: *every positive integer, different from* 1, *has at least one prime divisor.* Indeed, if the number n has no divisors except itself and 1, it is by definition prime and is its own prime divisor. If n has other divisors, then $n = ab$, where $a < n$, $b < n$. Consider a which again can be prime (and hence a prime divisor of n) or it has two factors: $a = a_1b_1$. Then $n = a_1(b_1b)$ where $a_1 < a$, i.e. a_1 is a divisor of n. Continuing this procedure we obtain a decreasing sequence of divisors of n: $a_r < \cdots < a_1 < n$. This sequence has to end somewhere. If it ends at a_r, then a_r is a prime divisor of n.

We are now able to prove an analog of Lemmas 2 and 4.

Theorem 5. *If the product of two positive integers is divisible by a prime, then at least one of them is divisible by that prime.*

Suppose that we want to prove the theorem for a prime p. We will prove it for all primes in the increasing order (as, in fact, we did it in the case of Lemma 2 for $p = 2$ and Lemma 4 for $p = 3$). Therefore, when we arrive at p, we can suppose the theorem has already been proved for all primes q smaller than p. Let $n_1 \cdot n_2$ be divisible by p and neither n_1 nor n_2 is divisible by p. Then

$$n_1 \cdot n_2 = pa. \tag{5}$$

Applying Theorem 4 to the pairs n_1, p and n_2, p, we get

$$n_1 = pt_1 + r_1, \quad n_2 = pt_2 + r_2,$$

where r_1 and r_2 are naturals less than p (and different from 0, for, otherwise, one of them would be divisible by p). Substituting in (5) and grouping the numbers divisible by p, we obtain

$$r_1r_2 = p(a - t_1r_2 - t_2r_1 - pt_1t_2),$$

or

$$r_1r_2 = pb, \quad b = a - t_1r_2 - t_2r_1 - pt_1t_2 \tag{6}$$

where, now, not alike in (5), $r_1 < p$ and $r_2 < p$. If $r_1 = 1$ and $r_2 = 1$, the contradiction $1 = pb$ is obtained. Let $r_1 > 1$. We know that r_1 has a prime factor q not greater than r_1 and, by that, less than p. Let $r_1 = qa_1$. The equality (6), now, implies

$$q(a_1r_1) = pb. \tag{7}$$

As already said, the theorem can be considered proved for all primes less than p and, in particular, for q. Being pb divisible by q, one of its factors must also be divisible by q. Since p is prime, b is divisible by q: $b = qb_1$. Substituting in (7) and after cancellation, we obtain

$$a_1r_2 = pb_1$$

and $a_1 < r_1$, $b_1 < b$. If $a_1 \neq 1$, proceeding again in the same way, another prime will be cancelled in the last equality. As the sequence a, a_1, ... of so obtained numbers is decreasing, we eventually come to an end, finishing with the number 1. Then, we have $r_2 = pb'$, which is impossible, being $r_2 < p$ (and $r_2 > 0$). Thus, the theorem has been proved.

You have noticed that the above reasoning is similar to the proofs of Lemmas 2 and 4: the statement reduces to the case when in (5), n_1 and n_2 (or better to say r_1 and r_2) are less than p. But here, consideration of all possible cases and direct checking are replaced by an elegant reasoning, by which, the theorem can be taken to be true for smaller values than p. (Euclid proved Theorem 5 somewhat differently. Most probably, the here presented reasoning belongs to Gauss.)

Now, to prove irrationality in general, no new ideas are needed.

Theorem 6. *If c is a positive integer which is not the square of any positive integer, then c is not the square of any rational number, i.e. $\sqrt{c}$ is irrational.*

We may again verify our statement, going from a positive integer to the bigger one, and thus, we can suppose the theorem has been proved for all smaller c's. In that case, we can assume that c is not divisible by the square of any positive integer greater than 1. Indeed, if $c = d^2 f$, $d > 1$, then $f < c$ and f is not the square of positive integer, since $f = g^2$ would imply $c = (dg)^2$ which contradicts the assumption of the theorem. Thus we can assume the theorem has already been proved for f, and accordingly, take that $\sqrt{f}$ is irrational. But then $\sqrt{c}$ is not rational, either. In fact, the equality $\sqrt{c} = \frac{n}{m}$, in view of $\sqrt{c} = d\sqrt{f}$, yields $\frac{n}{m} = d\sqrt{f}$, $\sqrt{f} = \frac{n}{dm}$, which would mean $\sqrt{f}$ is rational.

Now we proceed to the main part of the proof. Suppose $\sqrt{c}$ is rational and $\sqrt{c} = \frac{n}{m}$, where n and m are taken, as we already did it before, to be relatively prime. Then $m^2 c = n^2$. Let p be a prime factor of c. Put $c = pd$ and d is not divisible by p, otherwise c would be divisible by p^2 and now we consider the case when c is not divisible by a square. From $m^2 c = n^2$, it follows that n^2 is divisible by p and, according to Theorem 5, n is divisible by p. Let $n = pn_1$. Using $n = pn_1$ and $c = pd$ and substituting in the relation $m^2 c = n^2$, we get $m^2 d = pn_1^2$. Being m and n relatively prime and n divisible by p, m is not divisible by p. Then, according to Theorem 5, m^2 is not divisible by p, either. And, as we have seen it, d is not divisible by p because it would imply that c is divisible by p^2. Now, the equation $m^2 d = pn_1^2$ is contradictory to Theorem 5.

Notice that, in this section, we have more than once derived this or that property of positive integers taking them one after the other and, first, checking the property for $n = 1$ and, then, after supposing its validity for numbers less than n, we proved it for n.

Here we lean upon a statement which has to be considered as an axiom of arithmetic.

If a property of positive integers is valid for $n = 1$ (or $n = 2$) and if from its validity for all positive integers less than n, the validity for n follows, then the property is valid for all positive integers. This statement is called the Principle of Mathematical Induction or of Total Induction. Sometimes, instead of supposing the validity for all numbers less than n, only the validity for $n - 1$ is supposed. A statement which corresponds to the case $n = 1$ or $n = 2$, with which the reasoning starts (sometimes $n = 0$ is more convinient) is called the *basis of induction* and the statement corresponding to $n-1$, the *inductive hypothesis.* The Principle of Mathematical Induction is also used to produce a type of definitions, when a concept involving a positive integer n is defined by supposing that it has already been defined for $n-1$. For example, when we define an arithmetic progression using the property that each term is obtained from the preceding one by adding a constant d, called the common difference, then we have the type of a definition by induction. To express the definition symbolically, we write

$$a_n = a_{n-1} + d.$$

And to determine the entire progression we only need to know the initial term: a_1 or a_0.

In one of his treatise, French mathematician and physicist H. Poincaré considers the question: how is it possible that mathematics, which is founded on proofs containing syllogisms, that is statements expressed in a finite number of words, does lead to theorems related to infinite collections (for example, Theorem 6 holds for infinite set of numbers c; Theorem 2 asserts that $2n^2 \neq m^2$ for all positive integers n and m, the number of which is also infinite). Possibilities for it, Poincaré sees in the Principle of Mathematical Induction, which, in his words, "contains an infinite number of syllogisms condensed in a single formula".

Problems

1. Prove the irrationality of $\sqrt{5}$ by the same method used in the proofs of Theorems 2 and 3.

2. Prove that the number of positive integers divisible by m and less than n is equal to the integer part of the quotient, when n is divided by m.

3. Prove that if a positive integer c is not the cube of any positive integer, then $\sqrt[3]{c}$ is irrational.

4. Replace the reasoning, connected with succesive subtracting of the number m in the proof of Theorem 4, by a reference to the Principle of Mathematical Induction.

5. Using the Principle of Mathematical Induction, prove the formula

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

6. Using the Principle of Mathematical Induction, prove the inequality $n \leqslant 2^n$.

3. The prime factorization

In the previous section we have seen that every natural number has a prime divisor. Starting with this, we can get much more:

Theorem 7. *Every natural number greater than* 1 *can be expressed as a product of prime numbers.*

If the number p is itself prime, then the equality $p = p$ is considered as a prime factorization containing only one factor. If the number $n > 1$ is not prime, it has a prime divisor, different from itself: $n = p_1 \cdot n_1$ and (since by definition $p_1 \neq 1$), $n_1 < n$. Now we can apply the same reasoning to n_1 and continue. We obtain the factorization $n = p_1 \cdot \ldots \cdot p_k \cdot n_k$, where $p_1, \ldots, p_k$ are primes and the quotients n_k decrease: $n > n_1 > n_2 > \cdots$. Since our process must come to an end, we obtain $n_r = 1$ for some r and the factorization is $n = p_1 \cdot \ldots \cdot p_r$. Of course, the reader can easily formulate this proof in a more "scientific" way—using the method of mathematical induction.

The process used in the proof of Theorem 7 is not uniquely determined: if the number n has several prime factors, then any of them could be the first one. For example, 30 can be expressed first as $2 \cdot 15$ and then as $2 \cdot 3 \cdot 5$ and it is also possible to express it first as $3 \cdot 10$ and then as $3 \cdot 2 \cdot 5$. The fact that two resulting factorizations differ only by the order of their factors, was unpredictable. If for number 30 we could easily foresee all possibilities, is it so simple to convince oneself that the number

$$740037721 = 23623 \cdot 31327$$

has no other prime factorizations?

In school curricula it is usually assumed, as a self-evident fact, that every given natural number has only one prime factorization. However, this claim has to be proved, as the following example shows. Suppose that we know only even numbers and do not know how to use odd numbers. (It is possible that this is a reflection of the real historical situation, since in English the term "odd" has also the meaning "strange"). Repeating literally the definition of the prime number, we should call "prime" all even numbers which do not factorize into product of two *even* factors. For instance, the "prime" numbers would be 2, 6, 10, 14, 18, 22, 26, 30, ... Then a given number may have two different "prime" factorizations, for example

$$60 = 2 \cdot 30 = 6 \cdot 10.$$

It is also possible to find numbers with more different factorizations, such as

$$420 = 2 \cdot 210 = 6 \cdot 70 = 10 \cdot 42 = 14 \cdot 30.$$

Therefore, if the prime factorization is indeed unique, then in the proof of this statement we must use some properties which express that we are dealing with all natural numbers and not, say, with even numbers.

Since we are convinced that the uniqueness of the prime factorization is not self-evident, let us prove it.

Theorem 8. *Any two prime factorizations of a given natural number differ only by the order of their factors.*

The proof of the theorem is not really slef-evident, but all the difficulties have been already overcome in the proof of Theorem 5. From this theorem everything follows easily.

First, let us note an obvious generalization of Theorem 5.

If a product of any number of factors is divisible by a prime number p, then at least one of them is divisible by p.

Let

$$n_1 \cdot n_2 \cdot \ldots \cdot n_r = p \cdot a.$$

We shall prove our statement by induction on the number of factors r. When $r = 2$, it coincides with Theorem 5. If $r > 2$, write the equality in the form

$$n_1(n_2 \cdot \ldots \cdot n_r) = p \cdot a.$$

According to Theorem 5, either n_1 is divisible by p—and then the statement is proved—or $n_2 \cdot \ldots \cdot n_r$ is divisible by p—and then the statement is again true by the induction hypothesis.

We now prove Theorem 8. Suppose that a number n has two prime factorizations:

$$n = p_1 \cdot \ldots \cdot p_r = q_1 \cdot \ldots \cdot q_s. \tag{8}$$

We see that p_1 didvides the product $q_1 \cdot \ldots \cdot q_s$. By the generalization of Theorem 5, proved earlier, p_1 divides one of the numbers $q_1, \ldots, q_s$. But q_i is a prime number and its only prime divisor is itself. Hence, p_1 coincides with one of the q_i's. Changing their numeration we can take $p_1 = q_1$. Cancelling the equality (8) by p_1 we get

$$n' = \frac{n}{p_1} = p_2 \cdot \ldots \cdot p_r = q_2 \cdot \ldots \cdot q_s. \tag{9}$$

This is a statement concerning a smaller number n' and using mathematical induction we can take it to be true. Hence the number of factors in the two factorizations is the same, i.e. $r - 1 = s - 1$, implying $r = s$. Besides, the factors $q_2, \ldots, q_s$ can be written in such an order that $p_2 = q_2$, $p_3 = q_3, \ldots, p_r = q_r$. Since we have already established that $p_1 = q_1$, the theorem is proved.

The theorem we just proved can be found in Euclid. Although simple, it was always considered to be an abstract mathematical theorem. However, in the last two decades it found an unexpected practical application which we shall shortly comment. The application is connected with *coding*, i.e. writing an information in such a form that it cannot be understood by a person who does not know some additional information (the key of the code). Namely, it turns out that the problem of prime factorization of large numbers requests an enormous amount of operations; this problem is much more involved than the “inverse” problem—multiplication of prime numbers. For example, it is possible, though tedious, to multiply two prime numbers, each one having tens of digits (30 or 40 digits, say) in a day and to write down the result (which will have about 70 digits) in the evening. But factorization of this number into two primes would take more time, even if we use the best contemporary computer, than the time which elapsed since the formation of the Earth. Hence, a pair of large numbers p and q on one hand, and their product

$n = pq$ on the other, give, in view of Theorem 8, the same information, written in two different ways, but the transition from the pair p, q to the number $n = pq$ is easy, whereas the transition from n to the pair p, q is practically impossible. This is the underlying idea of coding; we omit technical descripiton.

In the prime factorization of a number, certain primes may appear several times, for insatnce, $90 = 2 \cdot 3 \cdot 3 \cdot 5$. We can group together the equal factors and write $90 = 2 \cdot 3^2 \cdot 5$. Hence, for each positive integer n we have the factorization

$$n = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot \ldots \cdot p_r^{\alpha_r}, \tag{10}$$

where all the primes $p_1, \ldots, p_r$ differ from one another and the exponents $\alpha_i \geqslant 1$. This factorization is said to be *canonical.* Of course, such a factorization is unique for every n.

Knowing the canonical factorization of a number n, we can find out whatever we want about its divisors. First, if the canonical factorization has the form (10), then it is obvious that the numbers

$$m = p_1^{\beta_1} \cdot \ldots \cdot p_r^{\beta_r}, \tag{11}$$

where $\beta_1 \leqslant \alpha_1$, $\beta_2 \leqslant \alpha_2, \ldots, \beta_r \leqslant \alpha_r$, are divisors of n, where β_i may have the value 0 (i.e. some of the p_i's which are divisors of n need not be divisors of m). Conversely, any divisor of n has the form (11). Indeed, if $n = mk$, then k is a divisor of n, i.e. it has the form (11): $k = p_1^{\gamma_1} \cdot \ldots \cdot p_r^{\gamma_r}$. Multiplying the canonical factorizations of m and k and grouping together the powers of equal primes, we have to arrive at the factorization (10), since this factorization is, by Theorem 8, unique. When two powers of a prime number are multiplied, their exponents add up which implies that $\beta_1 + \gamma_1 = \alpha_1$, i.e. $\beta_1 \leqslant \alpha_1$ and similarly $\beta_2 \leqslant \alpha_2, \ldots, \beta_r \leqslant \alpha_r$.

For example, we can find the sum of the divisors of n. We also take the number itself, n and also 1 to be its divisors. For instance, $n = 30$ has the divisors 1, 2, 3, 5, 6, 10, 15, 30 and their sum is 72. Consider first the simplest case when n is a power of a prime number: $n = p^\alpha$. Then its divisors are the numbers p^β where $0 \leqslant \beta \leqslant \alpha$, i.e. the numbers $1, p, p^2, \ldots, p^\alpha$. We therefore have to find the sum $1 + p + p^2 + \cdots + p^\alpha$. There is a general formula (which you may already know) which gives the sum of consecutive powers of a number:

$$s = 1 + a + \cdots + a^r.$$

The derivation of the formula is quite simple: we multiply both sides of the above equality by a:

$$sa = a + a^2 + \cdots + a^{r+1}.$$

We see that the expressions for s and sa consist of almost the same terms, but in s we have 1 which does not figure in sa, whereas in sa we have a^{r+1} which does not figure in s. Hence, after subtracting s from sa, all the terms cancel out, except those two:

$$sa - s = a^{r+1} - 1,$$

i.e. $s(a-1)=a^{r+1}-1$, and

$$s = 1 + a + a^2 + \cdots + a^r = \frac{a^{r+1}-1}{a-1}. \tag{12}$$

Since we have divided by $a-1$, we must suppose that $a \neq 1$.

Therefore, if $n = p^\alpha$ the sum of its divisors is $1 + p + \cdots + p^\alpha = \dfrac{p^{\alpha+1}-1}{p-1}$. Consider now the next case when n has two prime divisors p_1 and p_2. Its canonical factorization has the form $n = p_1^{\alpha_1} p_2^{\alpha_2}$. In view of formula (11), the divisors of n are $p_1^{\beta_1} p_2^{\beta_2}$, where $0 \leqslant \beta_1 \leqslant \alpha_1$, $0 \leqslant \beta_2 \leqslant \alpha_2$. Split them up into groups, one group for each value of β_2. So, for $\beta_2 = 0$ we obtain the divisors 1, p_1, p_1^2, ..., $p_1^{\alpha_1}$ whose sum is $\dfrac{p_1^{\alpha_1+1}-1}{p_1-1}$. For $\beta_2 = 1$ we obtain the group p_2, p_1p_2, $p_1^2p_2$, ..., $p_1^{\alpha_1}p_2$. In order to find the sum of those divisors, we notice that it is equal to $(1 + p_1 + p_1^2 + \cdots + p_1^{\alpha_1})p_2 = \dfrac{p_1^{\alpha_1+1}-1}{p_1-1}\,p_2$. Similarly, for any value of β_2 we obtain the sum $(1 + p_1 + p_1^2 + \cdots + p_1^{\alpha_1})p_2^{\beta_2} = \dfrac{p_1^{\alpha_1+1}-1}{p_1-1}\,p_2^{\beta_2}$. Hence, the total sum of divisors is

$$\begin{aligned}\frac{p_1^{\alpha_1+1}-1}{p_1-1} + \frac{p_1^{\alpha_1+1}-1}{p_1-1}\,p_2 + \frac{p_1^{\alpha_1+1}-1}{p_1-1}\,p_2^2 + \cdots + \frac{p_1^{\alpha_1+1}-1}{p_1-1}\,p_2^{\alpha_2} \\ = \frac{p_1^{\alpha_1+1}-1}{p_1-1}(1 + p_2 + p_2^2 + \cdots + p_2^{\alpha_2}).\end{aligned}$$

We evaluate the sum in the parentheses by another application of the formula (12). As a result we conclude that the sum of all divisors of $n = p_1^{\alpha_1}p_2^{\alpha_2}$ is $\dfrac{p_1^{\alpha_1+1}-1}{p_1-1} \cdot \dfrac{p_2^{\alpha_2+1}-1}{p_2-1}$.

We now pass on to the general case. Consider the product

$$S' = (1 + p_1 + p_1^2 + \cdots + p_1^{\alpha_1})(1 + p_2 + p_2^2 + \cdots + p_2^{\alpha_2}) \cdot \ldots \cdot (1 + p_r + p_r^2 + \cdots + p_r^{\alpha_r})$$

and remove the parentheses. How do we do that? If we have one pair of parentheses, i.e. an expression of the form $(a + b + \cdots)k$, we multiply each of the summands a, b, etc. by k and the result is the sum of ak, bk, etc. If we have two pairs of parentheses $(a_1 + b_1 + c_1 + \cdots)(a_2 + b_2 + c_2 + \cdots)$ we multiply each term from one parentheses by each term from the other and the result is the sum of all terms a_1a_2, a_1b_2, a_1c_2, b_1a_2, b_1b_2, etc. Finally, for any number of parentheses $(a_1 + b_1 + c_1 + \cdots)(a_2 + b_2 + c_2 + \cdots) \cdot \ldots \cdot (a_r + b_r + c_r + \cdots)$ we take one term from each, multiply them and then evaluate the sum of all such products. Apply this rule to our sum S'. The terms in the parentheses have the form $p_1^{\beta_1}$, $p_2^{\beta_2}$, ..., $p_r^{\beta_r}$ $(0 \leqslant \beta_i \leqslant \alpha_i)$. Multiplying them we get $p_1^{\beta_1}p_2^{\beta_2}\cdots p_r^{\beta_r}$, which is, in view of (11), a divisor of n, and according to Theorem 8, each one appears only once. Hence the sum S' is equal to the sum of the divisors of n. On the other hand, the

i-th parentheses, according to (12), is equal to $\dfrac{p_i^{\alpha_i+1}-1}{p_i-1}$, and the product of all parentheses is

$$S=\frac{p_1^{\alpha_1+1}-1}{p_1-1}\cdot\frac{p_2^{\alpha_2+1}-1}{p_2-1}\cdot\ldots\cdot\frac{p_r^{\alpha_r+1}-1}{p_r-1}.$$

This is the formula for the sum of all divisors. But we have also found the *number* of divisors. Indeed, in order to determine the number of divisors we have to replace each summand in the sum of divisors by 1. Returning to the previous proof, we see that it is enough to replace each summand in each parentheses of the product S' by 1. The first parentheses is then equal to α_1+1, the second to α_2+1, $\ldots$, the r-th to α_r+1. Hence, the number of divisors is $(\alpha_1+1)(\alpha_2+1)\cdots(\alpha_r+1)$. For example, the number of divisors of the number whose canonical factorization is $p^\alpha q^\beta$ is equal to $(\alpha+1)(\beta+1)$.

In the same way we can derive the formula for the sum of squares or cubes or generally k-th powers of the divisors of n. The reasoning is the same as the one applied for finding the sum of the divisors. Verify that the formula for the sum of k-th powers of all divisors of the number n with the canonical factorization (10) is

$$S=\frac{p_1^{k(\alpha_1+1)}-1}{p_1^k-1}\cdot\frac{p_2^{k(\alpha_2+1)}-1}{p_2^k-1}\cdot\ldots\cdot\frac{p_r^{k(\alpha_r+1)}-1}{p_r^k-1}. \tag{13}$$

We can also investigate common divisors of two positive integers m and n. Let their canonical factorizations be

$$n=p_1^{\alpha_1}\cdot\ldots\cdot p_r^{\alpha_r},\qquad m=p_1^{\beta_1}\cdot\ldots\cdot p_r^{\beta_r}, \tag{14}$$

where in any pair of numbers (α_i,β_i) one of them may have the value 0—this is the case when a prime number divides one of the numbers m, n, but not the other. Then on the basis of what we know about divisors we can say that the number k is a common factor of m and n if and only if it has the form

$$k=p_1^{\gamma_1}\cdot\ldots\cdot p_r^{\gamma_r},$$

where $\gamma_1\leqslant\alpha_1$, $\gamma_1\leqslant\beta_1$, $\gamma_2\leqslant\alpha_2$, $\gamma_2\leqslant\beta_2$, $\ldots$, $\gamma_r\leqslant\alpha_r$, $\gamma_r\leqslant\beta_r$. In other words if σ_i denotes the smaller of the numbers α_i, β_i, these conditions become $\gamma_1\leqslant\sigma_1$, $\gamma_2\leqslant\sigma_2$, $\ldots$, $\gamma_r\leqslant\sigma_r$. Put

$$d=p_1^{\sigma_1}\cdot\ldots\cdot p_r^{\sigma_r}. \tag{15}$$

The above reasoning proves that the following theorem is true.

Theorem 9. *For any two numbers with canonical factorization* (14), *the number d, defined by* (15), *divides both n and m, and any common factor of n and m divides d.*

The number d is called the *greatest common divisor* of n and m and is denoted by g. c. d.(n,m). It is clear that among all divisors of n and m, d is the greatest, but it is not obvious that all other common divisors divide it. This follows from

Theorem 8 (about the uniqueness of prime factorization). That is why we proved those properties which are usually given in school courses without proof.

As we said earlier, finding the prime factorization of a number is a very difficult task. Hence we give a different method for finding the greatest common divisor which does not use prime factorization—this method is often taught at schools. It is based on Theorem 4. Let n and m be two positive integers and let $n = mt + r$, $0 \leqslant r < m$ be the representation eastablished in Theorem 4.

Lemma 5. *If* $r \neq 0$, *then* $\mathrm{g.\,c.\,d.}(n,m) = \mathrm{g.\,c.\,d.}(m,r)$.

More than that: all common divisors of the pairs (n,m) and (m,r) are equal, and so are the greatest which are divisible by the others. Indeed, any common divisor d of numbers n and m is a divisor of m and of r, because $r = n - mt$, and a common divisor d' of m and r is a divisor of m and of n, because $n = mt + r$.

The transition from the pair (n,m) to the pair (m,r) is fruitful since $r < m$. We can now apply the same reasoning to the pair (m,r). Let $m = rt_1 + r_1$, $0 \leqslant r_1 < r$. If $r_1 \neq 0$, then $\mathrm{g.\,c.\,d.}(m,r) = \mathrm{g.\,c.\,d.}(r,r_1)$. We continue this process as long as we can. But the process ends when we get the remainder 0, for example $r_i = r_{i+1}t_{i+2} + 0$ $(r_{i+2} = 0)$. But then r_{i+1} divides r_i and clearly $\mathrm{g.\,c.\,d.}(r_i, r_{i+1}) = r_{i+1}$. Therefore, the last nonzero remainder in the process of dividing n by m, m by r, r by r_1, etc. is equal to $\mathrm{g.\,c.\,d.}(n,m)$. This method of finding the g. c. d. is called *Euclid's algorithm*, and it can be found in Euclid. For instance, in order to find $\mathrm{g.\,c.\,d.}(8891, 2329)$ we make the following divisions:

$$8891 = 2329 \cdot 3 + 1904; \qquad 2329 = 1904 \cdot 1 + 425;$$
$$1904 = 425 \cdot 4 + 204; \qquad 425 = 204 \cdot 2 + 17; \qquad 204 = 17 \cdot 12 + 0,$$

and conclude: $\mathrm{g.\,c.\,d.}(8891, 2329) = 17$.

The numbers n and m are said to be *relatively prime* if they have no common divisor other than 1. This means that $\mathrm{g.\,c.\,d.}(n,m) = 1$. Hence, using Euclid's algorithm we can find whether two numbers are relatively prime, without knowing their prime factorizations.

At the end of this chapter we return to the question with which we started: the question of irrationality. We shall prove a very wide generalization of our first assertion regarding the irrationality of $\sqrt{2}$. It is in connection with the concept to which we devote the next chapter and so this can be treated as a kind of introduction to that chapter.

An expression of the form ax^k, where a is a number, x is unknown and k a natural number or 0 (in which case we simply write a), is called *monomial.* The number k is its *degree*, and a is its *coefficient.* In general, we can consider monomials in several unknowns, such as ax^2y^8, but at the moment we are concerned only with monomials in one unknown. A sum of monomials is a *polynomial.* If a polynomial contains several monomials of the same degree, for example ax^k and bx^k, we can replace them by one monomial, namely $(a+b)x^k$. Having this in mind, we shall always assume that a polynomial contains only one member of a given degree k and we write it in the form a_kx^k; for $k = 0$ we simply have the number a_0. The greatest degree of all monomials which comprise a polynomial is the *degree* of the polynomial. For example, the degree of the polynomial $2x^3 - 3x + 7$ is 3 and its coefficients are $a_0 = 7$, $a_1 = -3$, $a_2 = 0$, $a_3 = 2$. Thus, a polynomial of degree n can be written in the form

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n,$$

where some of a_k's can be zero, but $a_n \neq 0$, because otherwise the degree of the polynomial would be less than n. The term a_0 is called the *constant term* of the polynomial, a_n is its *leading coefficient.* The equation $f(x) = 0$ is called an *algebraic equation* with one unknown. A number α is called its *root* if $f(\alpha) = 0$. A root of the equation $f(x) = 0$ is also called a *root of the polynomial* $f(x)$. Degree of the polynomial $f(x)$ is the *degree of the equation.* Obviuosly, the equations $f(x) = 0$ and $cf(x) = 0$, where c is a number, distinct from 0, are equivalent.

Now we shall treat such equations $f(x) = 0$ whose coefficients a_0, a_1, $\dots$, a_n are rational numbers, some of which may be equal to 0 or negative. If c is the common denominator of all, distinct from 0, coefficients, we can pass from the equation $f(x) = 0$ to the equation $cf(x) = 0$ having integer coefficients. In the sequel we shall treat only such equations. In this connection we shall have to deal with the divisibility of (not only nonnegative) integers. Recall that an integer a is, by definition, divisible by an integer b if $a = bc$ for some integer c.

Theorem 10. *Let $f(x)$ be a polynomial with integer coefficients and with leading coefficient equal to* 1. *If the equation $f(x) = 0$ has a rational root α, then α is an integer and it is a divisor of the constant term of the polynomial $f(x)$.*

Let us represent α in the form $\alpha = \pm\dfrac{a}{b}$, where the fraction $\dfrac{a}{b}$ is irreducible, i.e. positive integers a and b are relatively prime. By the condition, the polynomial $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_{n-1}x^{n-1} + x^n$ has integer coefficients a_i. Let us substitute α into the equation $f(x) = 0$. By the assumption,

$$a_0 + a_1\left(\pm\frac{a}{b}\right) + \dots + a_{n-1}\left(\pm\frac{a}{b}\right)^{n-1} + \left(\pm\frac{a}{b}\right)^n = 0. \tag{16}$$

Multiply the equation by b^n and transfer $(\pm a)^n$ to the right-hand side. All the terms remaining on the left-hand side will be divisible by b:

$$(a_0b^{n-1} + a_1(\pm a)b^{n-2} + \dots + a_{n-1}(\pm a)^{n-1}b^{n-2})b = (\pm 1)^{n-1}a^n.$$

We see that b divides a^n. If α were not an integer, b would be > 1. Let p be some of its prime divisors. Then it has to divide a^n, and by Theorem 5, p divides a, too. However, by the assumption, a and b are relatively prime and we have obtained a contradiction. Hence, $b = 1$ and $\alpha = a$.

In order to obtain the second assertion of the theorem, let us leave just a_0 on the right-hand side, and transfer all the other terms to the right-hand side (recall that $b = 1$). All the terms on the right-hand side are divisible by a:

$$a_0 = a(\mp a_1 - a_2(\pm a) - \dots - a_{n-1}(\pm a)^{n-2} - (\pm a)^{n-1}).$$

Obviously, it follows that a devides a_0.

Theorem 10 allows us to find rational roots of equations of the given form: in order to do that we have to list all the divisors of the constant term (with signs + and −) and check whether they are roots. For example, for the equation $x^5 - 13x + 6 = 0$ we have to check the numbers ±1, ±2, ±3, ±6. Only $x = -2$ is a root.

In such a way, all the roots of a polynomial $f(x)$ with integer coefficients and the leading coefficinet equal to 1 are irrational, except integer roots which are included as divisors of the constant term. That is just what we have proved in the beginning of this Chapter: firstly for $f(x) = x^2 - 2$ (Theorem 2), then for $f(x) = x^2 - 3$ (Theorem 3) and finally for $f(x) = x^2 - c$, where c is an integer (Theorem 6). Now we have obtained the widest generalization of all these assertions. It has a lot of other geometrical applications, besides Theorems 1, 2, 3, 6.

Consider, e.g., the equation

$$x^3 - 7x^2 + 14x - 7 = 0. \tag{17}$$

By Theorem 10, its integer roots can be just divisors of the number −7, i.e. one of the numbers 1, −1, 7, −7. Substitutions show that neither of these numbers satisfies the equation. We can conclude that the roots of the equation are irrational numbers. As a matter of fact, we do not know whether the equation (17) has roots at all. But, we shall show later that it has roots and even very interesting ones. One of its roots appears to be the square of the side of the regular heptagon, inscribed in the circle of radius 1. Moreover, the equation (17) has three roots, lying: between 0 and 1, between 2 and 3 and between 3 and 4. They are the squares of the *diagonals* of the regular heptagon, inscribed into the unit circle. Here by a diagonal we mean an *arbitrary* segment, joining two vertices of a polygon, so that sides are included as diagonals. The regular heptagon has three diagonals of different length—AB, AC and AD (fig. 5). In such a way, all these lengths are irrational numbers.

Fig. 5

Problems

1. Show that Theorem 5 is not valid if concepts of a number and a “prime” are understood as applied just to even numbers as has been discussed in the beginning of this section. Which part of the proof of Theorem 5 appears to be wrong in that case?

2. Prove that if integers m and n are relatively prime, then divisors of mn are obtained multiplying divisors of m by divisors of n, and each divisor of mn can be obtained in this way exactly once. Deduce that if $S(N)$ denote the sum of k-th powers of all divisors of N, m and n are relatively prime and $N = mn$, then $S(N) = S(m)S(n)$. Derive the formula (14) in that way.

3. A positive integer n is called *perfect* if it is equal to the sum of its proper divisors (i.e. the number itself is *excluded* from the set of its divisors). E.g. numbers 6 and 28 are perfect. Prove that if for some r the number $p = 2^r - 1$ is prime, then $2^{r-1}p$ is a perfect number (but recall that the formula on the sum of divisors S we have deduced, includes the number n itself). This proposition was already known to Euclid. Nearly 2000 years later Euler proved the inverse assertion: each even perfect number is of the form $2^{r-1}p$, where $p = 2^r - 1$ is a prime. Proof does not use any facts other than the ones presented previously, but is by no means easy. Try to rediscover this proof! By now, it is not known whether there exist *odd* perfect numbers.

4. If for two positive integers m and n there exist such integers a and b so that $ma + nb = 1$, then obviously m and n are relatively prime: each of their common divisors is divisible by 1. Prove the converse: for relatively prime numbers m and n there always exist integers a and b such that $ma + nb = 1$. Use the division algorithm and mathematical induction.

5. Using the result of problem 4 prove Lemmas 6 and 7 without using the theorem on uniqueness of prime factorization. Show that in such a way a new proof of this theorem (Theorem 8) can be obtained. This was just the way Euclid proved it.

6. Find integer values of a such that the polynomial $x^n + ax + 1$ has rational roots.

7. Let $f(x)$ be a polynomial with integer coefficients. Prove that if a reduced fraction $\alpha = \pm\frac{a}{b}$ is a root of the equation $f(x) = 0$, then b divides the leading coefficient and a divides the constant term. This is a generalization of Theorem 10 to the case of a polynomial with integer coefficients where the leading coefficient need not be equal to 1.

I. R. Shafarevich,
Russian Academy of Sciences,
Moscow, Russia

THE TEACHING OF MATHEMATICS
1999, Vol. II, 1, pp. 1–30

SELECTED CHAPTERS FROM ALGEBRA

I. R. Shafarevich

Abstract. This paper is the second part of the publication "Selected chapters of algebra", the first part being published in the previous volume of the Teaching of Mathematics, Vol. I (1998), 1-22.

AMS Subject Classification: 00 A 35

Key words and phrases: Polynomial, multiple roots and derivatives, binomial formula, Bernoulli's numbers.

CHAPTER II. POLYNOMIAL

1. Roots and divisibility of polynomials

In this chapter we shall be concerned with equations of the type $f(x)=0$, where f is a polynomial. We have already met with them at the end of the previous chapter. The equation $f(x)=0$ should be understood as the problem: find all the roots of the polynomial (or the equation). But it may happen that all the coefficients of the polynomial $f(x)$ are 0 and the equation $f(x)=0$ turn into an identity. We then write $f=0$ and in that case we agree that the degree of the polynomial f is not defined.

In order to add up two polynomials we simply add the corresponding members. Polynomial are multiplied using the bracket rules. If $f(x)=a_0+a_1x+\cdots+a_nx^n$, $g(x)=b_0+b_1x+\cdots+b_mx^m$, then $f(x)g(x)=(a_0+a_1x+\cdots+a_nx^n)(b_0+b_1x+\cdots+b_mx^m)$. Eliminating the brackets we obtain members $a_kb_lx^{k+l}$, where $0\leqslant k\leqslant n$, $0\leqslant l\leqslant m$. After that we group together similar members. As a result we obtain the polynomial $c_0+c_1x+c_2x^2+\cdots$ with coefficients

$$(1)\qquad c_0=a_0b_0,\quad c_1=a_0b_1+a_1b_0,\quad c_2=a_0b_2+a_1b_1+a_2b_0,\quad \ldots$$

The coefficient c_m is equal to the sum of all products a_kb_l, where $k+l=m$.

Polynomials share many properties with integers. The representation of a polynomial in the form $f(x)=a_0+a_1x+\cdots+a_nx^n$ can be considered to be an analog of the representation of a positive integer in the decimal (or some other) system. The degree of a polynomial has the role analogous to the absolute value

This paper is an English translation of: И. Р. Шафаревич, *Избраные главы алгебры*, Математическое образование, 1, 2, июль–сент. 1997, Москва, стр. 3–33. In the opinion of the editors, the paper merits wider circulation and we are thankful to the author for his kind permission to let us make this version.

of an integer. For example, if we prove a property of integers by induction on the absolute value, then in the proof of the analogous property of polynomials we use induction on the degree. Notice the following important property: the degree of the product of two polynomials is equal to the sum of their degrees. Indeed, let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ be polynomials of degree n and m, that is to say $a_n \neq 0$, $b_m \neq 0$. If we calculate the coefficients of $f(x)g(x)$ using (1), we obtain members of the form $a_kb_lx^{k+l}$ where $k + l \leqslant n + m$. Clearly, the greatest degree we get is $m + n$ and there is only one such member: $a_nb_mx^{n+m}$. It differs from zero, since $a_nb_m \neq 0$, and it cannot be cancelled with some other member, since it has the greatest degree. This property is analogous to the property $|xy| = |x||y|$ of the absolute value $|x|$ of a number x.

The theorem on division with a remainder for polynomials is formulated and proved almost in the same way as for positive integers (Theorem 4, Chapter I).

Theorem 1. *For any polynomials $f(x)$ and $g(x)$, where $g \neq 0$, there exist polynomials $h(x)$ and $r(x)$ such that*

$$f(x) = g(x)h(x) + r(x) \tag{2}$$

where either $r = 0$, or the degree of r is less than the degree of g. For given f and g, the polynomials h and r are uniquely determined.

If $f = 0$, then the representation (2) is obvious: $f = 0 \cdot g + 0$. Suppose that $f \neq 0$ and apply the method of mathematical induction on the degree of $f(x)$. Suppose that the degree of $f(x)$ is n and that the degree of $g(x)$ is m:

$$f(x) = a_0 + a_1x + \cdots + a_nx^n, \quad g(x) = b_0 + b_1x + \cdots + b_mx^m.$$

If $m > n$, the representation (2) has the form $f = 0 \cdot g + f$, with $h = 0$, $r = f$. If $m \leqslant n$, put $f_1 = f - \frac{a_n}{b_m}x^{n-m}g$ (remember that $b_m \neq 0$, since the degree of $g(x)$ is m). Clearly, in the polynomial f_1 the members having x^n cancel out (that is how we chose the coefficient $-\frac{a_n}{b_m}$), which means that its degree is less than n. Hence, we can take that the theorem is true for that polynomial and that it has the representation of the form (2): $f_1 = gh_1 + r$, where $r = 0$ or its degree is less than m. This implies $f = f_1 + \frac{a_n}{b_m}x^{n-m}g = \left(h_1 + \frac{a_n}{b_m}x^{n-m}\right)g + r$, and we have obtained the representation (2) with $h = h_1 + \frac{a_n}{b_m}x^{n-m}$. Let us prove that the representation (2) is unique. If $f = gk + s$ is another such representation (which means that $s = 0$ or its degree is less than m), then subtracting one from the other we get

$$g(h - k) + r - s = 0, \qquad g(h - k) = s - r.$$

If the polynomial $s - r$ is 0, then $s = r$ and $h = k$. If $s - r \neq 0$, then its degree is less than m and we arrive at a contradiction, since $s - r$ is equal to the polynomial $g(h-k)$, obtained by multiplying g by $h-k$, and hence its degree cannot be smaller than the degree of g which is m.

Now please read the proof of Theorem 4 of Chapter I in order to see that the above proof is perfectly analogous to it. On the other hand, if we do all the operations which should be done in the application of the mathematical induction (i.e. transition from f_1 to the polynomial f_2 with even smaller degree, etc. until we arrive at the remainder r whose degree is less than m), we obtain the rule for the so called "corner" division of polynomials used in schools. For example, if $f(x) = x^3 + 3x^2 - 2x + 5$, $g(x) = x^2 + 2x - 1$, then the division is done according to the scheme:

$$\begin{array}{rl} x^3+3x^2-2x+5 & |\,\underline{x^2+2x-1} \\ \underline{x^3+2x^2-\ \ x} & \ x+1 \\ x^2-\ \ x+5 & \\ \underline{x^2+2x-1} & \\ -3x+6 & \end{array}$$

This means that we choose the leading term of the polynomial $h(x)$ so that when it is multiplied by the leading term of $g(x)$ (that is x^2) the result is the leading term of $f(x)$ (that is x^3). Therefore the leading term of $h(x)$ is x. In the first row of the above table we have $f(x)$ and in the second $g(x)x$ (the product of $g(x)$ and the leading term of $h(x)$). Their difference is in the third row. We now choose the next term of the polynomial $h(x)$ so that when multiplied by the leading term of $g(x)$ (that is x^2) it becomes equal to the leading term of the polynomial in the third row (that is x^2). Hence the next term of $h(x)$ is 1. We now repeat the procedure. In the fifth row we obtain a polynomial of degree 1 (which is less than 2, the degree of $g(x)$) and so the procedure stops. We see that

$$x^3 + 3x^2 - 2x + 5 = (x^2 + 2x - 1)(x + 1) - 3x + 6.$$

As in the case of numbers, the representation (2) is called *division with remainder* of the polynomial $f(x)$ by the polynomial $g(x)$. Polynomial $h(x)$ is the *quotient* and polynomial $r(x)$ is the *remainder* in this division.

The division of polynomials is analogous to the division of numbers and it is even simpler, since when we add two terms of a certain degree we obtain a term of the same degree, and we do not transform into tens, hundreds, etc., as in the case of number division.

Repeating the reasoning given in Chapter I for numbers, we can apply Theorem 1 to find the greatest common divisor of two polynomials. In fact, using the notation of Theorem 1 we have the following analog of Lemma 5 from Chapter I: $\mathrm{g.c.d.}(f, g) = \mathrm{g.c.d.}(g, r)$; more precisely, the pairs (f, g) and (g, r) have the same common divisors. We can now use Euclid's algorithm as in Chapter I: divide with reaminder g by r: $g = rh_1 + r_1$, then r by r_1, and so on, to obtain the following sequence of polynomials: r, r_1, r_2, $\dots$, r_n whose degrees decrease. We stop at the moment when we obtain the polynomial $r_{k+1} = 0$, i.e. when $r_{k-1} = r_k h_k$. From the equalities $\mathrm{g.c.d.}(f, g) = \mathrm{g.c.d.}(r, r_1) = \cdots = \mathrm{g.c.d.}(r_{k-1}, r_k)$ we see that $\mathrm{g.c.d.}(f, g) = \mathrm{g.c.d.}(r_{k-1}, r_k)$. But since r_k is a divisor of r_{k-1}, then

g. c. d.$(r_{k-1}, r_k) = r_k$ and so g. c. d.$(f, g) = r_k$ — the last nonzero remainder in the Euclid's algorithm. It should be remarked that g. c. d.(f, g) is not uniquely defined, which was not the case when we dealt with positive integers. Namely, if $d(x)$ is a common divisor of $f(x)$ and $g(x)$, then so is $cd(x)$, where $c \neq 0$ is a number. Hence, g. c. d.(f, g) is defined up to a multiplicative constant.

Theorem 1 becomes particularly simple and useful in the case when $g(x)$ is a first degree polynomial. We can then write $g(x) = ax + b$, with $a \neq 0$. Since the properties of division by g are unaltered if g is multiplied by a number, we multiply $g(x)$ by a^{-1}, so that the coefficient of x is 1. Write $g(x)$ in the form $g(x) = x - \alpha$ (it will soon become evident why it is more convenient to write α with a minus). According to Theorem 1, for any polynomial $f(x)$ we have

$$f(x) = (x - \alpha)h(x) + r. \tag{3}$$

But in our case the degree of r is less than 1, i.e. it is 0: *r is a number.* Can we find this number without carrying out the division? It is very simple—it is enough to put $x = \alpha$ into (3). We get $r = f(\alpha)$, and so we can write (3) in the form

$$f(x) = (x - \alpha)h(x) + f(\alpha). \tag{4}$$

Polynomial $f(x)$ is divisible by $x - \alpha$ if and only if the remainder in the division is 0. But in view of (4) it is equal to $f(\alpha)$. We therefore obtain the following conclusion which is called *Bézout's theorem.*

Theorem 2. *Polynomial $f(x)$ is divisible by $x - \alpha$ if and only if α is its root.*

For example, the polynomial $x^n - 1$ has a root $x = 1$. Therefore, $x^n - 1$ is divisible by $x - 1$. We came across this division earlier: see formula (12) of Chapter I (where a is replaced by x and $r + 1$ by n).

In spite of its simple proof, Bézout's theorem connects two completely different notions: divisibility and roots, and hence it has important applications. For instance, what can be said about the *common roots* of polynomials f and g, i.e. about the solutions of the system of equations $f(x) = 0$, $g(x) = 0$? By Bézout's theorem the number α is their common root if f and g are divisible by $x - \alpha$. But then $x - \alpha$ divides g. c. d.(f, g) which can be found by Euclid's algorithm. If $d(x) =$ g. c. d.(f, g), then $x - \alpha$ divides $d(x)$, i.e. $d(\alpha) = 0$. Therefore, the question of common roots of f and g reduces to the question of the roots of d, which is, in general, a polynomial of much smaller degree. As an illustration, we shall determine the greatest common divisor of two second degree polynomials, which we write in the form $f(x) = x^2 + ax + b$ and $g(x) = x^2 + px + q$ (we can always reduce them to this form after multiplication by a number). We divide f by g according to the general rule:

$$\begin{array}{l|l}
x^2 + ax + b & \underline{\;x^2 + px + q\;} \\
\underline{x^2 + px + q} & \qquad 1 \\
(a - p)x + (b - q) &
\end{array}$$

The remainder is $r(x) = (a - p)x + (b - q)$ and we know that g. c. d.$(f, g) =$ g. c. d.(g, r). Consider the case $a = p$. If we also have $b = q$, then $f(x) = g(x)$ and the system of equations $f(x) = 0$, $g(x) = 0$ reduces to one equation $f(x) = 0$. If $b \neq q$, then $r(x)$ is a nonzero number and f and g have no common factor. Finally, if $a \neq p$, then $r(x)$ has unique root $\alpha = \dfrac{b-q}{p-a}$. We know that $\text{g.c.d.}(f,g) = \text{g.c.d.}(g,r)$ and it is enough to substitute this value of α into $g(x)$, in order to find out whether $g(x)$ is divisible by $x - \alpha$. We obtain the relation

$$\left(\frac{b-q}{p-a}\right)^2 + p\left(\frac{b-q}{p-a}\right) + q = 0,$$

or, multiplying it by nonzero number $(p-a)^2$, the equivalent relation

$$(b-q)^2 + p(b-q)(p-a) + q(p-a)^2 = 0. \tag{4'}$$

The second and the third member of this equality have common factor $p-a$. Taking it out we can rewrite the relation (4') in the form

$$(q-b)^2 + (p-a)(pb-aq) = 0.$$

The expression $D = (q-b)^2 + (p-a)(pb-aq)$ is called the *resultant* of the polynomials f and g. We have seen that the condition $D = 0$ is necessary and sufficient for the existence of a common factor of $f(x)$ and $g(x)$, provided that $p \neq a$. But for $p = a$ the condition $D = 0$ becomes $q = b$, and that is, as we have seen, equivalent to the existence of a common non-constant factor of $f(x)$ and $g(x)$. In general, it is possible to find for any two polynomials $f(x)$ and $g(x)$ of arbitrary degrees an expression made up from their coefficients, which equated to zero gives necessary and sufficient condition for the existence of their common nonconstant factor, but of course, the technicalities will be more difficult.

Another important application of Bézout's theorem is considered with the number of roots of a polynomial. Suppose that the polynomial $f(x)$ is not identically zero, i.e. $f \neq 0$, and that $f(x)$ has, besides α_1, another root α_2 such that $\alpha_2 \neq \alpha_1$. By Bézout's theorem, $f(x)$ is divisible by $x - \alpha_1$:

$$f(x) = (x-\alpha_1)f_1(x). \tag{5}$$

Put $x = \alpha_2$ into this equality. Since α_2 is also a root of $f(x)$, we have $f(\alpha_2) = 0$. This means that $(\alpha_2 - \alpha_1)f_1(\alpha_2) = 0$, and hence (since $\alpha_2 \neq \alpha_1$) that $f_1(\alpha_2) = 0$, i.e. that α_2 is a root of $f_1(x)$. Applying Bézout's theorem to the polynomial $f_1(x)$ we obtain the equality $f_1(x) = (x-\alpha_2)f_2(x)$ and substituting this into (5) we obtain

$$f(x) = (x-\alpha_1)(x-\alpha_2)f_2(x).$$

Suppose that the polynomial $f(x)$ has k different roots $\alpha_1, \alpha_2, \dots, \alpha_k$. Repeating our reasoning k times we see that $f(x)$ is divisible by $(x-\alpha_1)\cdots(x-\alpha_k)$:

$$f(x) = (x-\alpha_1)\cdots(x-\alpha_k)f_k(x). \tag{6}$$

Let n be the degree of $f(x)$. On the right-hand side of (6) we have a polynomial whose degree is not less than k, and on the left-hand side a polynomial of degree n. Hence $n \geqslant k$. We formulate this as follows.

Theorem 3. *The number of different roots of a polynomial which is not identically zero is not greater than its degree.*

Of course, if a polynomial is identically equal to 0, all the numbers are its roots. Theorem 3 was proved in the 17th century by philosopher and mathematician Descartes.

Using Theorem 3 we can answer the question we have avoided up to now: what is the meaning of the phrase *equality of polynomials*? One way is to write the polynomials in the form

$$f(x) = a_0 + a_1x + \cdots + a_nx^n, \quad g(x) = b_0 + b_1x + \cdots + b_mx^m$$

and to say that they are equal if all their coefficients are equal: $a_0 = b_0$, $a_1 = b_1$, etc. This is how we think of the equality $f = 0$—all the coefficients of f are zero. Another way to understand the term "equality" is as follows: polynomials $f(x)$ and $g(x)$ are equal if they take the same values when x is substituted by an arbitrary number, i.e. if $f(c) = g(c)$ for all c. We shall prove that these two meanings of the notion "equality" coincide. But, at first we have to make a distinction between them and in the first case we say that "$f(x)$ and $g(x)$ have equal coefficients" and in the second that "$f(x)$ and $g(x)$ have equal values for all values of x".

Evidently, if $f(x)$ and $g(x)$ have equal coefficients, then they have equal values for all x. The converse will be proved in a stronger form: we do not have to suppose that $f(x)$ and $g(x)$ coincide for all values x—it is enough to suppose that they have the same values for any $n+1$ values of x, where n is not less than the degrees of both polynomials.

Theorem 4. *Suppose that the degrees of the polynomials $f(x)$ and $g(x)$ are not greater than n and that they have same values for some $n+1$ different values of x. Then the coefficients of $f(x)$ and $g(x)$ are equal.*

Proof. Suppose that the polynomials $f(x)$ and $g(x)$ have equal values for $n+1$ values of x: $x = \alpha_1, \alpha_2, \ldots, \alpha_{n+1}$, i.e. that

$$f(\alpha_1) = g(\alpha_1), \quad f(\alpha_2) = g(\alpha_2), \quad \ldots, \quad f(\alpha_{n+1}) = g(\alpha_{n+1}).$$

Consider the polynomial $h(x) = f(x) - g(x)$ (here "=" denotes the equality of coefficients). We have seen that this implies that $h(\alpha) = f(\alpha) - g(\alpha)$ for any α, and, in particular, that $h(\alpha_1) = 0$, $h(\alpha_2) = 0, \ldots, h(\alpha_{n+1}) = 0$. But the degrees of f and g are not greater than n, and so the degree of h is not greater than n. This is a contradiction with Theorem 3, unless $h = 0$, i.e. unless all the coefficients of h are 0. This implies that the coefficients of f and g are equal.

From now on we can apply the term "equality" to polynomials without emphasizing in which one of the two senses.

Theorem 4 shows an interesting property of polynomials. Namely, if we know the values of a polynomial $f(x)$ of degree not greater than n for some $n+1$ values of the variable x, then its coefficients are uniquely determined, and so are its values for *all other* values of x. Notice that in the above sentence "coefficients are uniquely determined" means only that there *cannot be* two different polynomials with the given property. Hence, it is natural to raise the question of the *existence* of such a polynomial. Namely, suppose that we have $n+1$ different numbers x_1, x_2, . . . , x_{n+1} and also $n+1$ numbers y_1, y_2, . . . , y_{n+1}; is there a polynomial $f(x)$ of degree not greater than n such that $f(x_1)=y_1$, $f(x_2)=y_2$, . . . , $f(x_{n+1})=y_{n+1}$? Theorem 4 states only that if such a polynomial exists, then it is unique. The problem of constructing such a polynomial is called the *problem of interpolation*. It often appears in the processing of experimental data, when a quantity $f(x)$ is measured only for certain values $x=x_1$, $x=x_2$, . . . , $x=x_{n+1}$ and it is necessary to make a plausible aasumption about its values for other values of x. The data are given by the table

(7)

x	x_1	x_2	$\dots$	x_{n+1}
$f(x)$	y_1	y_2	$\dots$	y_{n+1}

One of the plausible assumptions would be to construct a polynomial of degree not greater than n such that $f(x_1)=y_1$, $f(x_2)=y_2$, . . . , $f(x_{n+1})=y_{n+1}$ and to assume that the required quantity is equal to $f(x)$ for all values of x. But does such a polynomial exist? We shall prove that it does and we shall find its formula. It is called the *interpolation polynomial* corresponding to table (7). In order to find the formula for the interpolation polynomial in the general case, we shall first consider the *simplest interpolation problem*, when in table (7) all the values y_1, y_2, $\dots$ y_{n+1} are 0, except one of them. Let $y_1=y_2=\cdots=y_{k-1}=y_{k+1}=\cdots=y_{n+1}=0$, so that the table becomes

x	x_1	x_2	$\dots$	x_{k-1}	x_k	x_{k+1}	$\dots$	x_{n+1}
$f(x)$	0	0	$\dots$	0	y_k	0	$\dots$	0

This means that the required interpolation polynomial $f_k(x)$ has the following roots: x_1, x_2, . . . , x_{k-1}, x_{k+1}, . . . , x_{n+1} (i.e. all the numbers x_1, ... x_{n+1} except x_k). But then it must be divisible by the product of the corresponding factors $x-x_i$. Since there are n factors, and since the degree of the polynomial cannot be greater than n, it can differ from this product only by a multiplicative constant. That is to say, we have to put

$$f_k(x)=c_k(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_{n+1}). \tag{8}$$

Conversely, any polynomial of that form satisfies the required conditions for all x_1, . . . , x_{n+1}, except perhaps for $x=x_k$. If it is to satisfy the condition for x_k, we put $x=x_k$ into (8) and from the obtained equality we get the value of c_k. Since $f_k(x_k)$ has to be equal to y_k, we obtain

$$c_k=\frac{y_k}{(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_{n+1})}$$

$$f_k(x)=c_k(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_{n+1}).$$

Using the auxiliary polynomial $F(x)=(x-x_1)\cdots(x-x_{n+1})$ of degree $n+1$ we can write the above formula in a different way. Namely, in that case the product

$(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_{n+1})$ is equal to $\dfrac{F(x)}{x-x_k}$. Putting $\dfrac{F(x)}{x-x_k} = F_k(x)$, we get

$$c_k = \frac{y_k}{F_k(x_k)}, \qquad f_k(x) = \frac{y_k}{F_k(x_k)}F_k(x). \tag{9}$$

Passing on to the general interpolation problem with the table (7) we only have to notice that its solution is the sum of all polynomials $f_k(x)$ which correspond to all the simplest interpolation problems:

$$f(x) = f_1(x) + f_2(x) + \cdots + f_{n+1}(x).$$

Indeed, if we put $x = x_k$ then all the members on the right-hand side become 0, except $f_k(x_k)$, and since $f_k(x)$ is the solution of the k-th simplest interpolation problem, we have $f_k(x_k) = y_k$. Finally, the degrees of $f_1(x), \ldots, f_{n+1}(x)$ are not greater than n and the same holds for their sum. We can write the obtained formula in the form

$$f(x) = \frac{y_1}{F_1(x_1)}F_1(x) + \frac{y_2}{F_2(x_2)}F_2(x) + \cdots + \frac{y_{n+1}}{F_{n+1}(x_{n+1})}F_{n+1}(x), \tag{10}$$

where $F_k(x) = \dfrac{F(x)}{x-x_k}$, $F(x) = (x-x_1)(x-x_2)\cdots(x-x_{n+1})$.

There is an unexpected identity which follows from the formula for the interpolation polynomial. Consider the interpolation problem corresponding to the table

x	x_1	x_2	$\ldots$	x_{n+1}
$f(x)$	x_1^k	x_2^k	$\ldots$	x_{n+1}^k

where k is a positive integer not greater than n or $k = 0$. On one hand it is evident that the polynomial $f(x) = x^k$ is the solution of this interpolation problem. On the other hand, we can write it down using formula (10) and we obtain that

$$x^k = \frac{x_1^k}{F_1(x_1)}F_1(x) + \frac{x_2^k}{F_2(x_2)}F_2(x) + \cdots + \frac{x_{n+1}^k}{F_{n+1}(x_{n+1})}F_{n+1}(x),$$

where $F(x) = (x-x_1)(x-x_2)\cdots(x-x_{n+1})$ and $F_i(x) = \dfrac{F(x)}{x-x_i}$. The polynomials $F_i(x)$ have degree n and the coefficient of x^n is 1. If $k < n$, then the polynomial on the right must also have degree less than n which means that all members with degree n must cancel out. In other words we have

$$\frac{x_1^k}{F_1(x_1)} + \frac{x_2^k}{F_2(x_2)} + \cdots + \frac{x_{n+1}^k}{F_{n+1}(x_{n+1})} = 0.$$

for $k < n$. If $k = n$, the coefficient of x^n must be equal to 1 and we have

$$\frac{x_1^n}{F_1(x_1)} + \frac{x_2^n}{F_2(x_2)} + \cdots + \frac{x_{n+1}^n}{F_{n+1}(x_{n+1})} = 1.$$

Notice that here $F(x) = (x-x_1)\cdots(x-x_{n+1})$, $F_k(x) = \dfrac{F(x)}{x-x_k}$, so that we have identities for arbitrary numbers $x_1, \ldots, x_{n+1}$.

Problems

1. Write down the last identities for $n = 1$ and 2, i.e. for $F(x) = (x-x_1)(x-x_2)$ and $F(x) = (x-x_1)(x-x_2)(x-x_3)$, and then verify them by direct calculation.

2. Divide $x^{n+1} - 1$ by $x - 1$ in order to obtain another derivation of the formula (12), Chapter I.

3. Divide with remainder $x^n - a$ by $x^m - b$. (Hint: the answer depends on the division of n by m.)

4. In deducing the formula (6) why was not possible to reason as follows: since $f(x)$ is divisible by all $x - \alpha_i$, it is divisible by their product? Verify that the assertion: if n is divisible by a and b then n is divisible by ab does not hold for numbers. Verify that it also does not hold for polynomials.

5. Prove that any polynomial can be written as the product of binomials $x-\alpha_i$ and a polynomial which has no roots. Prove that such a representation for a given polynomial is unique.

6. Let $F(x) = (x - x_1)\cdots(x - x_n)$ where $x_1, \dots, x_n$ are different from one another and let $f(x)$ be a polynomial of degree less than n. Prove that the fraction $\dfrac{f(x)}{F(x)}$ is equal to the sum of fractions of the form $\dfrac{a_k}{x - x_k}$, $k = 1, \dots, n$. Find the formula for a_k.

7. If $g(x)$ is a polynomial of degree less than n and if $x_1, \dots, x_{n+1}$ and the polynomial $F(x)$ have the same meaning as at the end of Section 1, prove that

$$\frac{g(x_1)}{F_1(x_1)} + \cdots + \frac{g(x_{n+1})}{F_{n+1}(x_{n+1})} = 0.$$

8. Let everything be the same as in Problem 7, except that the degree of $g(x)$ is n and the coefficient of x^n is a. Prove that

$$\frac{g(x_1)}{F_1(x_1)} + \cdots + \frac{g(x_{n+1})}{F_{n+1}(x_{n+1})} = a.$$

2. Multiple roots and derivative

The equation $x^2 - a = 0$ for $a > 0$ has two roots, given by $x = \sqrt{a}$ and $x = -\sqrt{a}$, where $\sqrt{a}$ is the arithmetic value of the square root of a. For $a = 0$ this gives two equal values. Similarly, the formula for the solution of an arbitrary quadratic equation sometimes gives two equal roots. Can similar situation happen for equations of arbitrary degree? At first the question itself seems to be meaningless. What does it mean that the equation $f(x) = 0$ has two *equal* roots? We can write any root of an equation on the paper as many times as we please, and all these numbers will be equal! But when we spoke of equal roots of the quadratic equation we used the formula for its solution. In the general case we shall also use some additional considerations in order to give a reasonable *definition* of what does it mean that the equation $f(x) = 0$ has two equal roots $x = \alpha$ and $x = \alpha$.

Such considerations are based on Bezout's theorem (Theorem 2). Let $x=\alpha$ be a rooot of $f(x)$. By Bezout's theorem $f(x)$ is divisible by $x-\alpha$ and we have $f(x)=(x-\alpha)g(x)$, where $g(x)$ is a polynomial whose degree is less than the degree of $f(x)$ by 1. If the polynomial $g(x)$ again has a root $x=\alpha$, we shall say that $f(x)$ *has two roots equal to* α. By Bezout's theorem, $g(x)$ can be written in the form $g(x)=(x-\alpha)h(x)$ and hence

$$f(x)=(x-\alpha)^2h(x). \tag{11}$$

We can say that in the representation (6) there are two factors $x-\alpha$. This is in accordance with the intuitive notion of what are two equal roots.

If in (11) $h(x)$ again has a root α, we shall say that $f(x)$ has *three roots* equal to α. In general, if $f(x)$ can be written in the form $f(x)=(x-\alpha)^r u(x)$, where $u(x)$ is a polynomial whose root is not α, we shall say that $f(x)$ has r *equal roots* α. If $r\geqslant 2$, then α is said to be a *multiple root*. Hence, α is a multiple root if $f(x)$ is divisible by $(x-\alpha)^2$. If the polynomial $f(x)$ has exactly k roots equal to α, we say that k is the *multiplicity* of the root α. Then $f(x)$ can be written in the form $f(x)=(x-\alpha)^k g(x)$, where α is not a root of $g(x)$, i.e. $g(\alpha)\neq 0$.

For example, suppose that $x=\alpha$ is a root of the quadratic equation $x^2+px+q=0$. Dividing x^2+px+q by $x-\alpha$ we get

$$\begin{array}{l|l}
x^2+px+q & x-\alpha \\ \cline{2-2}
\underline{x^2-\alpha x\qquad\quad} & x+p+\alpha \\
(p+\alpha)x+q & \\
\underline{(p+\alpha)x-\alpha(p+\alpha)} & \\
\qquad\qquad q+p\alpha+\alpha^2 &
\end{array}$$

i.e. $x^2+px+q=(x-\alpha)(x+p+\alpha)+(\alpha^2+p\alpha+q)$. Since α is a root of the equation $x^2+px+q=0$, we have $\alpha^2+p\alpha+q=0$, and so $x^2+px+q=(x-\alpha)(x+p+\alpha)$. By our definition, this equation has two roots equal to α if α is a root of $x+p+\alpha$, i.e. if $2\alpha+p=0$. Hence, $\alpha=-p/2$. Since $\alpha^2+p\alpha+q=0$, then putting $\alpha=-p/2$ we obtain that $-p^2/4+q=0$. This is the known condition which ensures that the equation $x^2+px+q=0$ has equal roots.

For the third order equation $x^3+ax^2+bx+c=0$ the calculation is only a little more involved. Divide x^3+ax^2+bx+c by $x-\alpha$:

$$\begin{array}{l|l}
x^3+ax^2+bx+c & x-\alpha \\ \cline{2-2}
\underline{x^3-\alpha x^2} & x^2+(a+\alpha)x+b+a\alpha+\alpha^2 \\
\quad (a+\alpha)x^2+bx+c & \\
\quad \underline{(a+\alpha)x^2-\alpha(a+\alpha)x} & \\
\qquad (b+a\alpha+\alpha^2)x+c & \\
\qquad \underline{(b+a\alpha+\alpha^2)x-\alpha(b+a\alpha+\alpha^2)} & \\
\qquad\qquad c+b\alpha+a\alpha^2+\alpha^3 &
\end{array}$$

Since by supposition $\alpha^3 + a\alpha^2 + b\alpha + c = 0$, then $x^3 + ax^2 + bx + c = (x - \alpha)(x^2 + (a + \alpha)x + b + a\alpha + \alpha^2)$. According to our definition the equation $x^3 + ax^2 + bx + c = 0$ has two roots equal to α if α is a root of the equation and also if α is a root of the polynomial $x^2 + (a + \alpha)x + b + a\alpha + \alpha^2$. In other words, $\alpha^2 + (a + \alpha)\alpha + b + a\alpha + \alpha^2 = 0$, i.e. $3\alpha^2 + 2a\alpha + b = 0$. We see that a multiple root of the equation $x^3 + ax^2 + bx + c = 0$ is the *common root* of the polynomials $x^3 + ax^2 + bx + c$ and $3x^2 + 2ax + b$. As we saw in Section 1, they are the roots of the polynomial $\text{g.c.d.}(x^3 + ax^2 + bx + c, 3x^2 + 2ax + b)$ and the greatest common divisor can be found by Euclid's algorithm.

We now apply the same reasoning to the polynomial $f(x) = a_0 + a_1x + \cdots + a_nx^n$ of arbitrary degree. When we divide $f(x)$ by $x - \alpha$ we obtain as the quotient a polynomial $g(x)$ of degree $n - 1$ whose coefficients depend on α and so we shall denote it by $g(x, \alpha)$. We know (formula (3)) that the remainder is $f(\alpha)$:

$$f(x) = (x - \alpha)g(x, \alpha) + f(\alpha). \tag{12}$$

Putting $x = \alpha$ into the polynomial $g(x, \alpha)$ we obtain the polynomial in α which is called the *derivative* of $f(x)$ and is denoted by $f'(\alpha)$. Hence, by definition,

$$f'(\alpha) = \frac{f(x) - f(\alpha)}{x - \alpha}(\alpha). \tag{13}$$

The above formula may cause some doubt, since after the substitution $x = \alpha$ both the numerator and the denominator in the expression $\dfrac{f(x) - f(\alpha)}{x - \alpha}$ become 0 and we get $\dfrac{0}{0}$. This formula therefore needs to be explained: we first (before substituting $x = \alpha$) divide the numerator by the denominator and we substitute $x = \alpha$ into their quotient which is a polynomial. For example, the meaning of the expression $\dfrac{x^2 - 1}{x - 1}(1)$ is: we first get $\dfrac{x^2 - 1}{x - 1} = x + 1$, and then $(x + 1)(1) = 2$.

Those of you who will continue to study mathematics will meet the derivative for other functions, such as $f(x) = \sin x$ or $f(x) = 2^x$. In essence they are defined by the same formula (13), but in general case it is more difficult to give the exact sense to the expression on right-hand side. In the case of polynomials everything is cleared by applying Bézout's theorem to the polynomial $f(x) - f(\alpha)$.

If α is a root of the polynomial $f(x)$ in (12), i.e. if $f(\alpha) = 0$, then we get $f(x) = (x - \alpha)g(x, \alpha)$ and by our definition α is a multiple root of $f(x)$ if α is a root of $g(x, \alpha)$, i.e. if $g(\alpha, \alpha) = 0$. But this means that $f'(\alpha) = 0$. We have proved the assertion:

Theorem 5. *A root of a polynomial $f(x)$ is multiple if and only if it is also a root of the derivative $f'(x)$.*

We see that a multiple root α is the *common root* of the polynomials $f(x)$ and $f'(x)$. In other words, α is a root of $\text{g.c.d.}(f(x), f'(x))$; the greatest common divisor can be found by Euclid's algorithm and it is, as a rule, a polynomial of much smaller degree.

We shall now carry out the division of $f(x)$ by $x-\alpha$, we shall find the polynomial $g(x,\alpha)$ in (12) and we shall find the explicite formula for the derivative of a polynomial.

We could make the usual division of $f(x)$ by $x-\alpha$ and find the quotient $g(x,\alpha)$ and the remainder $f(\alpha)$. But it is better to do it another way. Recall that $f(x)$ is the sum of the terms a_kx^k and hence $f(x)-f(\alpha)$ is the sum of the terms $a_k(x^k-\alpha^k)$. The polynomial $(x^k-\alpha^k)$ has a root $x=\alpha$ and by Bézout's theorem it is divisible by $x-\alpha$. We have noticed (after the formulation of Bézout's theorem) that we have already done this division earlier. True, only for $\alpha=1$, but the general case is easily reduced to it. We shall use formula (12) of Chapter I (where $r+1$ is replaced by k):

$$(x^k-1)=(x-1)(x^{k-1}+x^{k-2}+\cdots+x+1).$$

Replace x by x/α:

$$\left(\frac{x^k}{\alpha^k}-1\right)=\left(\frac{x}{\alpha}-1\right)\left(\frac{x^{k-1}}{\alpha^{k-1}}+\frac{x^{k-2}}{\alpha^{k-2}}+\cdots+\frac{x}{\alpha}+1\right).$$

Multiplying both sides of this equality by α^k we get

$$x^k-\alpha^k=(x-\alpha)(x^{k-1}+\alpha x^{k-2}+\cdots+\alpha^{k-2}x+\alpha^{k-1}). \tag{14}$$

This formula was obtained for $\alpha\neq 0$ (since we had x/α) but it is clearly true for $\alpha=0$ also.

Consider the polynomial $f(x)=a_0+a_1x+\cdots+a_nx^n$ and the difference $f(x)-f(\alpha)$. It is equal, as we saw, to the sum of the following terms $a_k(x^k-\alpha^k)$. Divide each such term by $x-\alpha$, using formula (14). We get

$$\frac{a_k(x^k-\alpha^k)}{x-\alpha}=a_k(x^{k-1}+\alpha x^{k-2}+\cdots+\alpha^{k-2}x+\alpha^{k-1}).$$

If we put $x=\alpha$ (into the right-hand side!) we obtain the term $ka_k\alpha^{k-1}$. Hence for the polynomial $g(x,\alpha)$ in (12) for $x=\alpha$ we get that $g(x,\alpha)(\alpha)=g(\alpha,\alpha)$ is the sum of terms $ka_k\alpha^{k-1}$, i.e. $a_1+2a_2\alpha+3a_3\alpha^2+\cdots+na_n\alpha^{n-1}$. In other words, we have deduced the formula for the derivative $f'(x)$ of the polynomial $f(x)=a_0+a_1x+\cdots+a_nx^n$:

$$f'(x)=a_1+2a_2x+\cdots+na_nx^{n-1}. \tag{15}$$

Compare this with what we obtained for polynomials of degree 2 and 3 and convince yourself that those were special cases of (15) for $n=2$ and $n=3$.

The derivative of a polynomial is important not only in connection with multiple roots; it has many other applications. We shall therefore prove the basic properties of the derivative. All the proofs follow from the definition, i.e. from (12).

a) *The derivative of a constant polynomial.* If $f(x)=a_0$ then, by definition, $f(x)=f(\alpha)$ and $g(x,\alpha)=0$. Hence $f'(\alpha)=0$, i.e. $f'(x)=0$.

b) *The derivative of a sum.* Let f_1 and f_2 be two polynomials and let $f = f_1 + f_2$. We have

$$\begin{aligned} f_1(x) &= f_1(\alpha) + (x-\alpha)g_1(x,\alpha), \\ f_2(x) &= f_2(\alpha) + (x-\alpha)g_2(x,\alpha) \end{aligned} \tag{16}$$

and therefore $f_1'(\alpha) = g_1(\alpha,\alpha)$, $f_2'(\alpha) = g_2(\alpha,\alpha)$. Adding the formulas (16) we get $f(x) = f(\alpha) + (x-\alpha)g(x,\alpha)$, where $g(x,\alpha) = g_1(x,\alpha) + g_2(x,\alpha)$. Therefore, $f'(\alpha) = g(\alpha,\alpha) = g_1(\alpha,\alpha) + g_2(\alpha,\alpha) = f_1'(\alpha) + f_2'(\alpha)$, i.e.

$$(f_1 + f_2)' = f_1' + f_2'.$$

Using induction on the number of summands we easily obtain

$$(f_1 + f_2 + \cdots + f_r)' = f_1' + f_2' + \cdots + f_r'.$$

c) *Multiplication by a number.* Let $f_1(x) = af(x)$. Then from the equalities $f(x) = f(\alpha) + (x-\alpha)g(x,\alpha)$ and $g(\alpha,\alpha) = f'(\alpha)$, multiplying by a we get

$$f_1(x) = af(x) = af(\alpha) + (x-\alpha)ag(x,\alpha),$$

i.e. $f_1(x) = f_1(\alpha) + (x-\alpha)ag(x,\alpha)$ and $f_1'(\alpha) = af'(\alpha)$:

$$(af)' = af'.$$

d) *The derivative of a product.* Let $f = f_1 f_2$. Multiplying the equalities (16) we get

$$f_1(x)f_2(x) = f_1(\alpha)f_2(\alpha) + (x-\alpha)g(x,\alpha),$$

where $g(x,\alpha) = g_1(x,\alpha)f_2(\alpha) + g_2(x,\alpha)f_1(\alpha) + (x-\alpha)g_1(x,\alpha)g_2(x,\alpha)$. Therefore $f(x) = f(\alpha) + (x-\alpha)g(x,\alpha)$ where $g(x,\alpha)$ is given above. Hence, $f'(\alpha) = g(\alpha,\alpha) = g_1(\alpha,\alpha)f_2(\alpha) + g_2(\alpha,\alpha)f_1(\alpha) = f_1'(\alpha)f_2(\alpha) + f_2'(\alpha)f_1(\alpha)$, i.e.

$$(f_1 f_2)' = f_1' f_2 + f_2' f_1. \tag{17}$$

If f_1 is a constant (polynomial of degree 0) then in view of a) from (17) we again get c).

By induction on the number of factors we obtain

$$(f_1 f_2 \cdots f_r)' = f_1' f_2 \cdots f_r + f_1 f_2' \cdots f_r + f_1 f_2 \cdots f_r' \tag{18}$$

(on the right-hand side in the product $f_1 \cdots f_r$ each factor is succesfully replaced by its derivative).

Indeed, according to (17):

$$(f_1 f_2 \cdots f_r)' = ((f_1 \cdots f_{r-1})f_r)' = (f_1 \cdots f_{r-1})' f_r + f_1 \cdots f_{r-1} f_r'.$$

Applying to $(f_1 \cdots f_{r-1})'$ the expression (18) which can be taken to be already proved, we obtain the required formula.

An important special case occurs when all the factors in (18) are equal:

$$(f^r)' = rf^{r-1}f'. \tag{19}$$

From the dfinition of the derivative it is easily verified that $x' = 1$. Hence, $(x^r)' = rx^{r-1}$. Combining the above rules, we can give a different proof of the explicite formula (15) for the derivative.

Return now to the question of multiple roots of polynomials. Suppose that α is a root of multiplicity k of $f(x)$. This means that $f(x) = (x-\alpha)^k g(x)$ where α in not a root of $g(x)$. According to (17) we have $f'(x) = ((x-\alpha)^k)'g(x) + (x-\alpha)^k g'(x)$, and according to (19) we have $((x-\alpha)^k)' = k(x-\alpha)^{k-1}$ (since $(x-\alpha)' = 1$, by (15)). Therefore, $f'(x) = k(x-\alpha)^{k-1}g(x) + (x-\alpha)^k g'(x) = (x-\alpha)^{k-1}p(-x)$, where $p(x) = kg(x) + (x-\alpha)g'(x)$. But α is not a root of $p(x)$: $p(\alpha) = kg(\alpha) \neq 0$. Consider the polynomials $d(x) = \text{g.c.d.}(f(x), f'(x))$ and $\varphi(x) = \dfrac{f(x)}{d(x)}$ (since $d(x)$ is a divisor of $f(x)$, $\varphi(x)$ is a polynomial). The polynomial $d(x)$ is divisible by $(x-\alpha)^{k-1}$ since $f(x)$ and $f'(x)$ are divisible by $(x-\alpha)^{k-1}$. But $d(x)$ is not divisible by $(x-\alpha)^k$, since $p(\alpha) \neq 0$ which means that $p(x)$ is not divisible by $x-\alpha$. We conclude that $\varphi(x)$ is divisible only by $x-\alpha$ (and no higher power, e.g. $(x-\alpha)^2$, etc). Since $\varphi(x)$ is defined independently from the root (namely $\varphi(x) = \dfrac{f(x)}{\text{g.c.d.}(f(x), f'(x))}$) the above conclusion is true for all the roots of $f(x)$, and we see that $\varphi(x)$ has the same roots as $f(x)$, but none of them is multiple. In view of this, we can always reduce a question regarding the roots of a polynomial to the case when the polynomial has no multiple roots.

Notice that we have implicitly met with the derivative in connection with the formula for the interpolation polynomial. Indeed, let $F(x) = (x-x_1)\dots(x-x_{n+1})$. From (14) we see that $(x-x_i)' = 1$. Therefore, formula (18) gives:

$$F'(x) = (x-x_2)\cdots(x-x_{n+1}) + (x-x_1)(x-x_3)\cdots(x-x_{n+1}) + \\ + \cdots + (x-x_1)(x-x_2)\cdots(x-x_n).$$

If we use the notation $F_k(x) = \dfrac{F(x)}{x-x_k}$ from Section I, then $F'(x) = F_1(x) + \cdots + F_{n+1}(x)$. Substituting now for x one of the values $x = x_k$, since all $F_i(x)$ for $i \neq k$ contain the factor $x - x_k$, we see that $F_i(x_k) = 0$. Therefore $F'(x_k) = F_k(x_k)$ and the formula (10) can be written in the form

$$f(x) = \frac{y_1}{F'(x_1)}F_1(x) + \frac{y_2}{F'(x_2)}F_2(x) + \frac{y_{n+1}}{F'(x_{n+1})}F_{n+1}(x).$$

Problems

1. Polynomial $x^{2n} - 2x^n + 1$ clearly has a root $x = 1$, and by Bézout's theorem it is divisible by $x - 1$. Find the quotient.

2. For which values of a, b does the polynomial $x^n + ax^{n-1} + b$ have a multiple root? Find this root.

3. For which values of a, b does the polynomial $x^3 + ax + b$ have a multiple root?

4. Prove that the polynomial $x^n + ax^m + b$ cannot have nonzero root of multiplicity 3 or more.

5. The derivative of the polynomial $f'(x)$ is called the *second derivative* of the polynomial $f(x)$ and is denoted by $f''(x)$. Find the formula for $(f_1f_2)''$, analogous to (17), but which will be, of course, somewhat more complicated.

6. Prove that the derivative of a polynomial is identically equal to 0 if and only if the polynomial is constant (i.e. when its degree is 0).

7. Prove that for a polynomial $f(x)$ there exists a polynomial $g(x)$ such that $g'(x) = f(x)$ and that all such polynomials $g(x)$ (for a given $f(x)$) can differ from each other only in the constant term.

8. Prove that the number of roots of a polynomial cannot be greater than its degree, each root being counted k times if k is its multiplicity.

3. The binomial formula

In this section we shall be concerned with an important formula which expresses the polynomial $(1+x)^n$ in the usual form $a_0 + a_1x + \cdots + a_nx^n$. In order to find the formula we have to multiply out all the factors $(1+x)(1+x)\cdots(1+x)$. Working out these brackets we shall obtain terms of the form x^k, but such terms will appear several times, and by grouping them together we shall arrive at the required formula. For instance, if $n = 2$, it is well known that

$$(1+x)^2 = (1+x)(1+x) = 1(1+x) + x(1+x) = 1 + x + x + x^2 = 1 + 2x + x^2.$$

For $n = 3$ the formula is also probably known. If not, it is easily obtained when the formula for $(1+x)^2$ is multiplied by $1+x$:

$$\begin{aligned}(1+x)^3 &= (1+x)^2(1+x) = (1+2x+x^2)(1+x)\\ &= (1+2x+x^2) + (1+2x+x^2)x = 1 + 3x + 3x^2 + x^3.\end{aligned}$$

The coefficient a_k of x^k in the polynomial $(1+x)^n$ depends on the index k, but also on the degree n. In order to indicate this dependence on n and k, we denote this coefficien by C_n^k. Therefore, C_n^k are *by definition* the coefficients in the formula

$$(1+x)^n = C_n^0 + C_n^1x + C_n^2x^2 + \cdots + C_n^nx^n. \tag{20}$$

For example, $C_2^0 = 1$, $C_2^1 = 2$, $C_2^2 = 1$; $C_3^0 = 1$, $C_3^1 = 3$, $C_3^2 = 3$, $C_3^3 = 1$. The coefficients C_n^k are called *binomial coefficients.* Our aim is to write them in an explicit form. Notice that some of them are easy to find. It is clear that when we multiply all the x's by one another in the product $(1+x)^n$, we get x^n, which means that the leading term of the polynomial $(1+x)^n$ is x^n, i.e.

$$C_n^n = 1. \tag{21}$$

Similarly, multiplying the constant terms (values for $x = 0$) in the product $(1+x)^n$ we see that the constant term of the polunomial $(1+x)^n$ is 1, i.e.

$$C_n^0 = 1. \tag{22}$$

In the general case consider the derivatives of both sides of (20). On the left, according to (19), we get $n(1+x)^{n-1}$, since $(1+x)'=1$, by (15). We evaluate the derivative on the right using (15). We obtain

$$n(1+x)^{n-1} = C_n^1 + 2C_n^2 x + \cdots + kC_n^k x^{k-1} + \cdots + nC_n^n x^{n-1}.$$

But we can apply (20) for $n-1$ to the left-hand side of the above equality. The coefficient of x^{k-1} will be nC_{n-1}^{k-1} on the left and kC_n^k on the right. Therefore, $kC_n^k = nC_{n-1}^{k-1}$, or

$$C_n^k = \frac{n}{k}\,C_{n-1}^{k-1},$$

i.e. the coefficient C_n^k can be expressed in terms of the coefficient C_{n-1}^{k-1} with smaller indices. Applying this formula to C_{n-1}^{k-1} we get $C_n^k = \dfrac{n(n-1)}{k(k-1)}C_{n-2}^{k-2}$, and repeating the process r times we obtain the formula

$$C_n^k = \frac{n(n-1)\cdots(n-r+1)}{k(k-1)\cdots(k-r+1)}C_{n-r}^{k-r}$$

(we take away from n in the numerator and from k in the denominator r consecutive values: $0,1,\dots,r-1$). Finally, let $r=k$. Since we know that $C_m^0=1$ for any m, we obtain the formula for C_n^k:

$$C_n^k = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}. \tag{23}$$

This is the formula we looked for.

Formula (20) with the explicit expression (23) for the binomial coefficients C_n^k is called the binomial formula (or "Newton's binomial").

The binomial formula has a large number of applications and it is useful to have the coefficients (23) written in various forms. In the denominator we have the product of all positive integers from 1 to k. The product of the form $1\cdot 2\cdot \ldots \cdot m$ is called *m factorial* and denoted by $m!$. In the numerator we have the product of all positive integers from n to $n-k+1$. If we multiply it by the product of the numbers from $n-k$ to 1 (i.e. by $(n-k)!$) we obtain $n!$. Therefore, multiplying the numerator and the denominator in (23) by $(n-k)!$, we get

$$C_n^k = \frac{n!}{k!\,(n-k)!}, \tag{24}$$

and this implies that

$$C_n^k = C_n^{n-k}. \tag{25}$$

Notice that in formulas (23) and (24) it is not immediately clear that the denominator divides the numerator, although we know that this is so having in mind the meaning of the coefficients C_n^k in the formula (20). We can express the fact that the expression on the right-hand side of (23) is an integer, by simply saying that the *product of any k consecutive integers is divisible by $k!$*. We shall see later that the fact that right-hand sides of (23) and (24) are integers implies some interesting properties of prime numbers.

We now establish some important properties of the coefficients C_n^k. The first one follows from the obvious equality $(1+x)^n = (1+x)^{n-1}(1+x)$ after expanding $(1+x)^n$ and $(1+x)^{n-1}$ on the basis of (20). We obtain

$$C_n^0 + C_n^1 x + C_n^2 x^2 + \cdots + C_n^n x^n = (C_{n-1}^0 + C_{n-1}^1 x + \cdots + C_{n-1}^{n-1} x^{n-1})(1+x).$$

The coefficient of x^k on the left is C_n^k and on the right is obtained from the sum of the terms $C_{n-1}^k x^k \cdot 1$ and $C_{n-1}^{k-1} x^{k-1} \cdot x$, i.e. it is $C_{n-1}^k + C_{n-1}^{k-1}$. Therefore

$$C_n^k = C_{n-1}^k + C_{n-1}^{k-1}. \tag{26}$$

This is a very useful formula for evaluating coefficients C_n^k by means of the coefficients of index $n-1$. In order to get a better visual representation, we write the coefficients C_n^k in the form of a triangle, where C_n^k are in the n-th row. Using the formulas (21) and (22), which say that at the beginning and at the end of each row is 1, the triangle has the form

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & \cdots & & \cdots & & \cdots & & \cdots & & & \\
 & & 1 & & C_{n-1}^2 & & \cdots & & C_{n-1}^{n-2} & & 1 & & \\
 & 1 & & C_n^1 & & C_n^2 & & \cdots & & C_n^{n-1} & & 1 & \\
\cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots
\end{array}$$

Formula (26) shows that each binomial coefficient C_n^k is equal to the sum of the coefficients which are situated above on the left and right of it. Taking the first two rows as given, we easily obtain for the subsequent coefficients:

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
\cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots
\end{array}$$

This triangle is called “Pascal’s triangle”.

The second property is obtained by putting $x = 1$ into the formula (20) which defines the binomial coefficients. On the left we get 2^n and on the right the sum of all binomial coefficients C_n^k for $k = 0, 1, \ldots, n$. Therefore, *the sum of all numbers from the n-th row of Pascal's triangle is equal to 2^n.*

Finally, consider two neighbouring members from one row: C_n^{k-1} and C_n^k. According to (24) we have $C_n^k = \dfrac{n!}{k!\,(n-k)!}$, $C_n^{k-1} = \dfrac{n!}{(k-1)!\,(n-k+1)!}$. Since $k! = (k-1)!\,k$, $(n-k+1)! = (n-k)!\,(n-k+1)$, we get

$$C_n^k = \frac{n-k+1}{k}\, C_n^{k-1}.$$

It is evident that $\dfrac{n-k+1}{k} > 1$ when $n-k+1 > k$, i.e. $k < \dfrac{n+1}{2}$ and in that case $C_n^k > C_n^{k-1}$. Conversely, if $k > \dfrac{n+1}{2}$, we obtain $C_n^k < C_n^{k-1}$. Therefore, *the numbers in one row of Pascal's triangle increase up to the middle of the row, and after that they decrease.* If n is even, then in the middle of the row we have the greatest number $C_n^{n/2}$, and if n is odd, then there are two neighbouring equal greatest numbers: $C_n^{(n-1)/2}$ and $C_n^{(n+1)/2}$. In that case, for $k = \dfrac{n+1}{2}$ we have $C_n^k = C_n^{k-1}$.

The formula (20), where the binomial coefficients are defined by (23) can be written in a somewhat more general form. In order to do that, put $x = b/a$ and multiply both sides of (20) by a^n. We obtain the formula

$$(a+b)^n = C_n^0 a^n + C_n^1 a^{n-1} b + C_n^2 a^{n-2} b^2 + \cdots + C_n^n b^n. \tag{27}$$

This formula was proved for $a \neq 0$ (since we divided by a), but it is obviously true also for $a = 0$. It is also called the binomial formula.

We shall now consider some consequences of the binomial formula and their applications. As a rule, the simpler a result is, the more applications it has. So, in the binomial formula we often use the values of the first coefficients. We already know that the first coefficient C_n^0 is 1. The next one C_n^1, according to (23) is n. Notice that in view of (25) it follows that $C_n^n = 1$ (which we already know) and that $C_n^{n-1} = n$. Hence,

$$(a+b)^n = a^n + na^{n-1}b + \cdots + nab^{n-1} + b^n.$$

This can be applied to equations. We write an equation of order n in the form $a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + a_n x^n = 0$. The fact that its degree is n means that $a_n \neq 0$ and we can divide the equation by a_n to obtain an equivalent equation in which $a_n = 1$. In further text we shall suppose that this has been done and we write the equation in the form $f(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} + x^n = 0$. We shall now make another transformation of this equation into an equivalent equation. In order to do so, put $x = y + c$, where y is the new variable and c is a number. Substituting into our equation this value of x, from each term $a_m x^m$ we obtain the term $a_m (y+c)^m$ which can be, by the binomial formula, written as a polynomial in y, and then we collect together corresponding terms. As a result we obtain a new polynomial in y which we denote by $g(y) = f(y+c)$. Since y is expressed in terms of x: $y = x - c$, the equations $f(x) = 0$ and $g(y) = 0$ are equivalent: to the root $x = \alpha$ of $f(x) = 0$ corresponds the root $y = \alpha - c$ of $g(y) = 0$, and to the root $y = \beta$ of this equation corresponds the root $x = \beta + c$ of the equation $f(x) = 0$. Let us examine how the coefficients change in this transformation. First of all, the degree of the equation $g(y) = 0$ is n and the coefficient of its leading term is 1. This follows from the fact that when $a_m(y+c)^m$ is expanded by the binomial formula, it gives rise to terms in y with degrees $\leqslant m$. Therefore, the term of degree n can only be obtained from $(y + c)^n$ and (again by the binomial formula) it is equal to y^n. Let us look at the term of degree $n - 1$. It can be obtained from the term $(y + c)^n$ and the term $a_{n-1}(y + c)^{n-1}$. From the last one we get $a_{n-1}y^{n-1}$, and in $(y + c)^n$ we have to take the *second* term in the binomial expansion. As we know it is equal to $ny^{n-1}c$. Hence, the term of degree $n - 1$ in the polynomial $g(y) = f(y + c)$ has the form $(a_{n-1} + nc)y^{n-1}$.

This can be used to simplify the equation by chosing c so that the term of degree $n - 1$ vanishes: we put $a_{n-1} + nc = 0$, i.e. $c = -a_{n-1}/n$. We proved the following

Theorem 6. *The substitution $x = y-a_{n-1}/n$ transforms the equation $f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n = 0$ into equivalent equation $g(y) = 0$ of degree n whose coefficient of the leading term is 1 and which has no term of degree $n - 1$.*

Notice that Theorem 6 gives the formula for the solutions of second degree equations. Indeed, the polynomial $g(y)$ has the form $y^2 + b_2$ and its roots are therefore $y = \pm\sqrt{-b_2}$. Make the substitution indicated in Theorem 6, evaluate b_2 and the roots of $f(x)$ and verify that in this way we obtain the standard formula for the solutions of a quadratic equation. In the case of polynomials of arbitrary degree we only get a certain simplification, which is sometimes useful. For example, we see that any third degree equation is equivalent to an equation of the form $x^3 + ax + b = 0$.

At the end we shall apply the binomial formula to the evaluation of the sums of powers of integers. We shall be concerned with the sums

$$S_m(n) = 0^m + 1^m + 2^m + \cdots + n^m \tag{28}$$

of m-th powers of all nonnegative integers not greater than n. You probably know the formula $S_1(n) = \dfrac{n(n+1)}{2}$ (see Problem 5 in Section 1 of Chapter I). We start with a few remarks on the evaluation of sums in general. Let $a_0, a_1, a_2, \ldots, a_n, \ldots$ be an arbitrary infinite sequence of numbers, and consider the following sequence of their sums: a_0, $a_0 + a_1$, $a_0 + a_1 + a_2$, $\ldots$, $a_0 + a_1 + a_2 + \cdots + a_n$, $\ldots$ Denote the first sequence by the letter a; its $(n + 1)$-st term is a_n (it is more convenient to write the $(n + 1)$-st term and not the n-th, and to start the sequence with a_0). The above sequence of sums will be denoted by Sa, and its $(n + 1)$-st term is

$$(Sa)_n = a_0 + a_1 + a_2 + \cdots + a_n, \qquad n = 0, 1, 2, \ldots$$

For example, if $a_n = n^m$, $n = 0, 1, 2, \ldots$, then Sa is the sequence of sums $S_m(n)$. Clearly, if we know the sequence Sa we can find the sequence a. Namely, by substracting the n-th from the $(n+1)$-st term of Sa, we obtain a_n. Indeed, let

$$b_n = (Sa)_n = a_0 + a_1 + \cdots + a_n, \tag{29}$$

$$b_{n-1} = (Sa)_{n-1} = a_0 + a_1 + \cdots + a_{n-1}, \tag{30}$$

and subtracting (30) from (29) we get $b_n - b_{n-1} = a_n$. We introduce another important construction. Together with an arbitrary sequence b_0, b_1, b_2, ... , b_n, ... consider the sequence b_0, $b_1 - b_0$, $b_2 - b_1$, ... , $b_{n+1} - b_n$, ... If the first sequence is denoted by b, then the second is denoted by Δb. Its $(n+1)$-st term is

$$(\Delta b)_0 = b_0, \quad (\Delta b)_n = b_n - b_{n-1}, \qquad n = 1, 2, \ldots$$

The established connection between the sequences a and Sa can be expressed by the formula $\Delta Sa = a$. It turns out that there is a formula completely symmetrical to this one, namely both equalities

$$\Delta Sa = a, \qquad S\Delta b = b \tag{31}$$

are true. It can be said that the operations S and Δ applied to sequences are inverse to each other. We have already established the first formula. In order to prove the second, write the equalities which define the numbers $a_k = (\Delta b)_k$ for $k = 0, 1, \ldots, n-1$:

$$\begin{aligned} a_0 &= b_0 \\ a_1 &= b_1 - b_0 \\ a_2 &= b_2 - b_1 \\ &\ldots \\ a_n &= b_n - b_{n-1} \end{aligned}$$

and add them up. On the left we obtain $a_0 + \cdots + a_n$, i.e. $(Sa)_n$, and on the right all the numbers cancel out, except b_n in the last formula, so that we get $(Sa)_n = b_n$, that is to say the second formula (31).

The above relations are useful, since it is often simpler not to evaluate a sum directly, i.e. not to find the sequence Sa directly, but instead to find a sequnce b such that $\Delta b = a$, and from the second relation (31) to obtain $Sa = b$.

This idea will now be applied to the sums (28). We have seen that $S_m(n) = (Sa)_n$, where $a_n = n^m$. How can we write the sequence a, $a_n = n^m$ in the form $a = \Delta b$? This follows from the following assertion.

Theorem 7. *For any polynomial $f(x)$ of degree m there exists a unique polynomial $g(x)$ of degree $m+1$ such that*

$$g(x) - g(x-1) = f(x) \tag{32}$$

and the constant term of $g(x)$ is 0.

The uniqueness of the polynomial $g(x)$ with the given property is easily shown. Let $g_1(x)$ be another polynomial such that $g_1(x) - g_1(x-1) = f(x)$ and whose constant term is 0. Subtract (32) from the last equality. Putting $g_1(x) - g(x) = g_2(x)$, we see that $g_2(x) - g_2(x-1) = 0$ and the constant term of $g_2(x)$ is 0, i.e. $g_2(0) = 0$. For $x = 1$, the above equality gives $g_2(1) = 0$. Putting $x = 2$ we get $g_2(2) = g_2(1) = 0, \ldots$ and by induction that $g_2(n) = 0$ for all positive integers n. In other words, all positive integers are roots of $g_2(x)$. According to Theorem 3 this is possible only if $g_2 = 0$, which means that $g = g_1$.

The existence of the polynomial g will be proved by induction on m, the degree of $f(x)$. For $m = 0$, the polynomial f is a constant a and we see that $g(x) = ax$ satisfies (32). Suppose that the assertion is true for polynomials f of degree less than m. Let $a_m x^m$ be the leading term of f. Choose the number a so that the leading term of the polynomial $ax^{m+1} - a(x-1)^{m+1}$ is equal to the leading term $a_m x^m$ of f. In order to do this, apply the binomial formula

$$(x-1)^{m+1} = x^{m+1} - (m+1)x^m + \cdots ,$$

where the dots stand for terms of degree less than m. This implies

$$x^{m+1} - (x-1)^{m+1} = (m+1)x^m + \cdots .$$

Clearly we have

$$a = \frac{a_m}{m+1}. \tag{33}$$

Then, in the difference $f(x) - \dfrac{a_m}{m+1}(x^{m+1} - (x-1)^{m+1})$ the terms of degree m cancel out and this difference will have degree less than m. Denoting this polynomial by $h(x)$, by the induction hypothesis we can take that there is a polynomial g_1 of degree less than $m+1$ and with zero constant term such that $h(x) = g_1(x) - g_1(x-1)$, i.e.

$$f(x) - \frac{a_m}{m+1}(x^{m+1} - (x-1)^{m+1}) = g_1(x) - g_1(x-1).$$

The above equality can be written in the form $f(x) = g(x) - g(x-1)$, where

$$g(x) = \frac{a_m}{m+1}x^{m+1} + g_1(x),$$

and the theorem is proved. Of course, in practical construction we do not apply induction, but we repeat the same procedure of subtraction to the polynomial $h(x)$, and so on until we arrive at a polynomial of degree 0.

Return now to the evaluation of the sum $S_m(n)$. We have seen that this sum is equal to b_n, where b is such that $\Delta b = a$, $a_n = n^m$. Apply Theorem 7 to the polynomial x^m. We obtain the polynomial $g(x)$ of degree $m+1$ such that

$$g(x) - g(x-1) = x^m$$

and the constant term of $g(x)$ is 0. Putting $x = n$ into the above equality, we see that the sequence $b_n = g(n)$ for $n \geqslant 1$ and $b_0 = g(0) = 0$ satisfies the condition $\Delta b = a$, i.e. $Sa = b$. Therefore we have proved the following

Theorem 8. *The sums $S_m(n)$ can be expressed in the form $g_m(n)$ where g_m is the polynomial of degree $m+1$ such that $g_m(x)-g_m(x-1)=x^m$ and its constant term is* 0.

Notice that the proof of Theorem 7 provides us with a method of constructing the polynomial $g_m(x)$ for any m. For instance, let $m=2$. By analogy with sequences, we denote the polynomial $g(x)-g(x-1)$ by Δg, i.e. we put $(\Delta g)(x)=g(x)-g(x-1)$. We first have to find the monomial ax^3 so that the leading term of $\Delta(ax^3)$ is equal to x^2. In view of (33), $a=\frac{1}{3}$ (in this case $m=2$, $a_2=1$). By the binomial formula, $\Delta\left(\frac{1}{3}x^2\right)=\frac{1}{3}x^3-\frac{1}{3}(x-1)^3=x^2-x+\frac{1}{3}$ and $x^2-\Delta\left(\frac{1}{3}x^3\right)=x-\frac{1}{3}$. Now we have to find the monomial bx^2 so that the leading coefficient of $\Delta(bx^2)$ is equal to x. In view of (33), $b=\frac{1}{2}$ (in this case $m=1$, $a_1=1$) and by the binomial formula $\Delta\left(\frac{1}{2}x^2\right)=\frac{1}{2}x^2-\frac{1}{2}(x-1)^2=x-\frac{1}{2}$, and $x^2-\Delta\left(\frac{1}{3}x^3\right)-\Delta\left(\frac{1}{2}x^2\right)=-\frac{1}{3}+\frac{1}{2}=\frac{1}{6}$. Finally, $\frac{1}{6}=\Delta\left(\frac{1}{6}x\right)=\frac{1}{6}x-\frac{1}{6}(x-1)$. At the end we get that $x^2=\Delta\left(\frac{1}{3}x^3+\frac{1}{2}x^2+\frac{1}{6}x\right)$ and so $g(x)=\frac{1}{3}x^3+\frac{1}{2}x^2+\frac{1}{6}x=\frac{(2x^2+3x+1)x}{6}=\frac{(2x+1)(x+1)x}{6}$. Therefore $S_2(n)=\frac{(2n+1)(n+1)n}{6}$.

We conclude with two more remarks.

Remark 1. The obtained formula for the sum $S_m(n)$ can be summarized as follows. For each m there exists the unique polynomial $g_m(x)$ with constant term 0 such that $g_m(x)-g_m(x-1)=x^m$. The method of its construction is contained in the proof of Theorem 7. Its degree is $m+1$. The formula for $S_m(n)$ is: $S_m(n)=g_m(n)$. Hence the question reduces to the investigation of the important polynomials $g_m(x)$. They are called *Bernoulli's polynomials.* In the Appendix we shall give a much more explixit expression for these polynomials, using an important sequence of rational numbers, called *Bernoulli's numbers.*

Remark 2. (Historical) The introduced operations S and Δ which transform the sequences a and b into Sa and Δb are very similar to the fundamental operations of Analysis which define for a function $f(x)$ (but not for every function!) the *indefinite integral* $\int f\,dx$, and for a function g its *derivative* g'. Our operations S and Δ are elementary analogs of the operations $\int f\,dx$ and g'. Sums and differences are also present in the definitions of the integral and the derivative, but in a more complicated way (in our definition of the derivative of a polynomial differences were also present—see formula (13)). As in the case of S and Δ, the operations of forming the derivative and the integral are inverse to each other. As in our case, the evaluation of the derivative is simpler than the evaluation of the integral, and the integral of a function $f(x)$ is mainly evaluated by finding a function whose derivative is equal to $f(x)$.

However, the operations for sequences and functions are not only analogous;

Fig. 1

their connection is deeper. Evaluating the integral of a function $f(x)$ is equivalent to evaluating the area of the surface bounded by the graph of that function, the x-axis and by two vertical lines starting at $x = a$ and $x = b$ (Fig. 1).

Of course, we shall not prove this, as we have not defined the integral, but we shall show, on a simple example, how such an area can be evaluated, and its connection with the problems we considered earlier.

We shall try to determine the area bounded by the parabola which is the graph of the function $y = x^2$, by the x-axis and by the line $x = 1$ (Fig. 2).

Fig. 2

In order to do that, divide the segment between 0 and 1 into a large number n of equal parts with coordinates $0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1$ and evaluate the corresponding values $0, \left(\frac{1}{n}\right)^2, \left(\frac{2}{n}\right)^2, \dots, \left(\frac{n-1}{n}\right)^2, 1$ of the function $y = x^2$. Construct the rectangles whose bases are segments from $\frac{i}{n}$ to $\frac{i+1}{n}$ and whose heights are $\left(\frac{i}{n}\right)^2$. The polygon made up from these rectangles is contained in that part "under the parabola" whose area we wish to determine and by "looking at the picture" we see that if n is very great, then the area s_n of this polygon differs very little from the area of the part under the parabola (we cannot be more precise, since we have not precisely defined what *area* is). The area of the polygon is the sum of the areas of the rectangles which make it up. The area of the i-th rectangle is equal to the product of its basis $\frac{1}{n}$ and its height $\left(\frac{i}{n}\right)^2$, i.e. it is $\frac{i^2}{n^3}$. Therefore, the area s_n of the polygon is

$$s_n = \frac{0^2}{n^3} + \frac{1^2}{n^3} + \frac{2^2}{n^3} + \cdots + \frac{(n-1)^2}{n^3} = \frac{S_2(n-1)}{n^3}.$$

We have already found that $S_2(n) = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n$, and so (replacing n by $n-1$ in the formula for $S_2(n)$) we get

$$s_n = \frac{1}{3} - \frac{1}{2} \cdot \frac{1}{n} + \frac{1}{6} \cdot \frac{1}{n^2}.$$

It is clear that as n becomes greater and greater, then the terms $-\frac{1}{2} \cdot \frac{1}{n}$ and $\frac{1}{6} \cdot \frac{1}{n^2}$ become smaller and smaller, and the area of the polygon approaches $\frac{1}{3}$. Hence, this is the area of the figure bounded by the parabola.

We have presented here the lines of thought followed, in principle, by Archimedes (3rd century B.C.) who was the first to solve this problem. (Archimedes devised a rather artificial method which allowed him to use the sum of a geometrical progression, instead of the sum $S_2(n)$. But he knew the formula for $S_2(n)$ and used it for the evaluation of other areas and volumes).

Mathematicians of the new period were obsessed by the dream to "surpass the ancients" (that is to say, the Ancient Greek mathematicians) and Archimedes was considered to be the most important of them. They were therefore very much interested in solving the problem considered above for the function $y = x^n$, where n is greater than 2. It seems that the first to obtain the solution was French mathematician Fermat (17th century) who used practically the same method we outlined above (it was later somewhat simplified). At that time the mentioned connection between the integral and the derivative was not known and the integral (i.e. the area) was calculated directly from the definition. It was later discovered that (to use contemporary terminology) the operations of forming derivatives and integrals are inverse to each other. This was established by Newton's teacher Barrow. (Newton worked together with Barrow when he studied at the university, and later on took over Barrow's chair when the latter decided to take orders). Systematic evaluation of the integral of a function f by finding a function g such that the derivative of g is f was initiated by Newton. After that the calculation of integrals and areas by the method we exposed became unnecessary. Nowadays students of higher classes can easily find the integral of x^m for any m without caluclating the sum $S_m(n)$.

In this way, if in Chapter I we moved in the circle of ideas of Ancient Greek mathematicians (Pythagoras, Theaetetus, Euclid), in this chapter we have encountered the ideas of the mathematicians from the new period (17th century).

Problems

1. Notice that the area s_n of the polygon which we calculated at the end of the section is *less* than the area of the given figure, bounded by the parabola $y = x^2$, since the polygon is situated inside that figure. Construct the polygon made up from rectangles whose bases are segments from $\frac{i}{n}$ to $\frac{i+1}{n}$ and whose heights are $\left(\frac{i+1}{n}\right)^2$ which *contains* the given figure. Its area s'_n will therefore be *greater* than the area of the figure. Calculate the area s'_n and prove that as n increases, it approaches 1/3. This gives a more convincing (i.e. more "strict") proof of the fact that the required area is 1/3.

2. Try to solve the analogous problem for the "m-th degree parabola", given by the equation $y = x^m$. Verify that in order to obtain the result it is not necessary to know the Bernoulli's polynomials $g_m(x)$ completely, but that is enough to know the coefficient of the leading term $a_{m+1}x^{m+1}$. Prove that $a_{m+1} = \frac{1}{m+1}$ and hence find the area of the figure bounded by the parabola whose equation is $y = x^m$, by the x-axis and by the line $x = 1$.

3. Prove that the area of the figure bounded by the parabola $y = x^m$, x-axis and the line $x = a$ is equal to $\frac{1}{m+1}a^{m+1}$. Notice that the derivative of the polynomial $\frac{1}{m+1}x^{m+1}$ is x^m. This is indeed an instance of Barrow's theorem that integration and finding derivatives are operations inverse to each other.

4. Prove that the sum of the binomial coefficients with even upper indices $C_n^0 + C_n^2 + \cdots$ and with odd indices $C_n^1 + C_n^3 + \cdots$ are equal and find their mutual value.

5. Find the relation between binomial coefficients which expresses that $(1+x)^n(1+x)^m = (1+x)^{n+m}$. For $n = m$ deduce the formula for the sum of the squares of binomial coefficients.

6. If p is a prime number, prove that all binomial coefficients C_p^k for $k \neq 0, p$, are divisible by p. Deduce from this that $2^p - 2$ is divisible by p. Prove that for any integer n, the number $n^p - n$ is divisible by p. This theorem was first proved by Fermat.

7. What can be said about the sequence a if all the terms of the sequence Δa are equal? What does formula (31) give in this case?

8. Find the sum $S_3(n)$ and verify that $S_3(n) = (S_1(n))^2$.

9. Let a be any sequence $a_0, a_1, a_2, \ldots$ Apply the operation Δ once more to the sequence Δa. The obtained sequence $\Delta(\Delta a)$ will be denoted by $\Delta^2 a$. Define $\Delta^k a$ by induction as $\Delta(\Delta^{k-1}a)$. When can we solve the so-called "infinite interpolation problem", that is to say when is there a polynomial $f(x)$ of degree not greater than m such that $f(n) = a_n$ for $n = 0, 1, 2, \ldots$? Prove that a necessary and sufficient condition is given by $(\Delta^{m+1}a)_n = 0$ for $n \geqslant m$. This condition shows that if we write the sequence a, and under it the sequence whose terms are differences of the two terms above, and so on:

$$\begin{array}{ccccccccccc}
a_0 & & a_1 & & a_2 & & \dots & & a_n & & a_{n+1} \\
& a_1 - a_0 & & a_2 - a_1 & & \dots & & \dots & & a_{n+1} - a_n & \dots \\
& & \dots & & \dots & & \dots & & \dots & & \dots
\end{array}$$

then in the $(m+1)$-st row we only have zeros.

Is there a polynomial $f(x)$ such that $f(n) = 2^n$ for all positive integers n?

10. Prove that if $a_n = q^n$, then $(\Delta a)_n = a_{n-1}(q-1)$. Use this to give another proof of the formula (12) from Chapter I.

11. Let $m_1 < m_2 < \cdots < m_{n+1}$ be positive integers and let $f(x)$ be a polynomial of degree n, where the coefficient of x^n is 1. Prove that at least one of the numbers $f(m_k)$ is not less than $n!/2^n$.

Hint. Use the result of Problem 8 of Section 1. Notice (with the notation of Section 1) that $F_k(m_k) \geqslant k!\,(n-k)!$ and use some known relations for binomial coefficients.

12. Apply the formula (12) from Chapter I to the sum $1 + (1+x) + (1+x)^2 + \cdots + (1+x)^n$. Equating the coefficients of terms with equal degrees on the left and right, find the formula for the sum

$$C_k^k + C_{k+1}^k + \cdots + C_n^k.$$

APPENDIX[1]

Bernoulli's polynomials and numbers

In Section 3 we showed that the values of the sums of powers of consecutive positive integers, i.e. the sums $S_m(n)$ coincide with the values $g_m(n)$ of Bernoulli's polynomials $g_m(x)$, which have the properties

$$\begin{aligned} &1)\ g_m(x) - g_m(x-1) = x^m, \\ &2)\ \text{the constant term of } g_m(x) \text{ is } 0. \end{aligned} \tag{1}$$

For any m there is only one polynomial of degree $m+1$ with these properties.

We have given a method for constructing the polynomials $g_m(x)$. However, we would like to have a more explicit formula for these polynomials. In order to achieve this, we shall follow the same path we took in deducing the binomial formula. Namely, we shall first find the derivative of both sides of (1). But we first have to see how to find the derivative $f(x-1)'$ of the polynomial $f(x-1)$.

Lemma 1. $f(x-1)' = f'(x-1)$.

[1] Starting with Chapter II, each chapter will have an Appendix. In these appendices we shall only use those facts which have been exposed earlier, but the text will be somewhat harder than the basic text. This means that we shall apply the same arguments as before, but in proving a theorem we shall have to keep in mind a larger number of them. The level of these texts approaches the level of a simple professional mathematical book.

At first sight this equality may seem obvious, but it is not really so. The equality means that when in the polynomial $f(x)$ we first substitute $x-1$ for x, then write it as the sum of powers of x, and then find its derivative, the obtained result is the same as when in the derivative $f'(x)$ we substitute $x-1$ for x.

The proof follows directly from the definition of the derivative of a polynomial, i.e. from (11). Let

$$f(x)-f(\alpha)=(x-\alpha)g(x,\alpha).$$

Substituting $x-1$ for x and $\alpha-1$ for α in this equality, we get

$$f(x-1)-f(\alpha-1)=(x-\alpha)g(x-1,\alpha-1).$$

By definition we have $f(\alpha-1)'=g(\alpha-1,\alpha-1)$ and $f'(\alpha)=g(\alpha,\alpha)$. Hence, $f(\alpha-1)'=f'(\alpha-1)$, which was to be proved.

Lemma 1 could be proved by the use of formulas (16)–(19) and reduction to monomials (verify this).

We can now find the derivatives of both sides of (1). Having in mind Lemma 1 and the rule (13) for derivatives, we obtain

$$g_m'(x)-g_m'(x-1)=mx^{m-1}.$$

On the other hand, replacing m by $m-1$ in (1) we get

$$g_{m-1}(x)-g_{m-1}(x-1)=x^{m-1}.$$

Multiply the second equality by m and subtract it from the first. Putting $h_m=mg_{m-1}-g_m'$ we find that

$$h_m(x)=h_m(x-1).$$

But this implies that the polynomial h_m is constant (of degree 0). Indeed, putting in this equality $x=1$, 2, etc. we obtain $h_m(0)=h_m(1)=h_m(2)=\cdots$. In other words the polynomial $h_m(x)$ and the constant $h_m(0)$ have equal values for all positive integers x, and in view of Theorem 4 they must be equal: $h_m(x)=h_m(0)$. (We have already met with this kind of reasoning at the beginning of the proof of Theorem 7.) Hence, the polynomial $h_m(x)$ is equal to a constant which we shall denote by $-\alpha_m$. Having in mind the definition of $h_m(x)$ we obtain the relation

$$g_m'=mg_{m-1}+\alpha_m. \tag{2}$$

As in the derivation of the binomial formula, we now write the polynomial $g_m(x)$ as the sum of powers of x. As before, the lower index indicates the polynomial in question, and upper index corresponds to the degree of x. The coefficients are denoted by A_m^k and $g_m(x)$ has the form

$$g_m(x)=A_m^1x+A_m^2x^2+\cdots+A_m^kx^k+\cdots+A_m^{m+1}x^{m+1}.$$

(Remember that the constant term of g_m is 0.) Write down the derivative of $g_m(x)$, using the formula (13):

$$g_m(x)'=A_m^1+2A_m^2x+\cdots+kA_m^kx^{k-1}+\cdots+(m+1)A_m^{m+1}x^m.$$

On the other hand, write down the analogous formula for $g_{m-1}(x)$ (replacing m by $m-1$):

$$g_{m-1}(x) + A^1_{m-1}x + A^2_{m-1}x^2 + \cdots + A^k_{m-1}x^k + \cdots + A^m_{m-1}x^m,$$

and substitute these two formulas into (2). Equating the coefficients of x^{k-1}, we find:

$$kA^k_m = mA^{k-1}_{m-1} \quad \text{for } k \geqslant 2, \tag{3}$$

$$A^1_m = \alpha_m \quad \text{for } k = 1. \tag{4}$$

(Notice that in the above formulas there is no α_0; we only have α_k where $k \geqslant 1$.) We have obtained the formula similar to the formula for the binomial coefficients C^k_m, the difference being that formula (3) holds only for $k \geqslant 2$, and for $k = 1$ it is replaced by (4).

Again, we continue to follow the case of binomial coefficients. We have: $A^k_m = \dfrac{m}{k} A^{k-1}_{m-1}$. Applying this formula to A^{k-1}_{m-1} we get: $A^k_m = \dfrac{m(m-1)}{k(k-1)} A^{k-2}_{m-2}$. Continuing this procedure, after $k-1$ steps we find

$$A^k_m = \frac{m(m-1)\cdots(m-k+2)}{k(k-1)\cdots 2} A^1_{m-k+1} = \frac{m(m-1)\cdots(m-k+2)}{k(k-1)\cdots 2} \alpha_{m-k+1}$$

(for A^1_{m-k+1} we use formula (4)).

The coefficient of α_{m-k+1} very much resembles the binomial coefficient. It differs from C^k_m (see formula (21)) in so much that the numerator does not have the last factor $m-k+1$ (and the denominator does not have the last factor 1, but this has no effect on the product). However, in the formula for C^k_{m+1} the product in the numerator ends with $m-k+1$, but it begins with $m+1$, which is not present here. Hence, we can write the coefficient of α_{m-k+1} in the form $\dfrac{1}{m+1}C^k_{m+1}$, and the formula for A^k_m becomes:

$$A^k_m = \frac{1}{m+1}C^k_{m+1}\alpha_{m+1-k}.$$

(We write α_{m-k+1} as α_{m+1-k} so that the factors look more similar.)

In this way we have obtained the following formula for the polynomials $g_m(x)$:

$$g_m(x) = \frac{1}{m+1}(C^1_{m+1}\alpha_m x + C^2_{m+1}\alpha_{m-1}x^2 + \cdots + \\ + C^k_{m+1}\alpha_{m+1-k}x^k + \cdots + C^{m+1}_{m+1}\alpha_0 x^{m+1}). \tag{5}$$

The obtained formula resembles the binomial formula. Suppose that we have a new variable a and expand the binomial $(x+a)^{m+1}$. We then obtain the same terms as in the brackets in the above formula (5), except that a^k is replaced by α_k and it has no term corresponding to $C^0_{m+1}\alpha_{m+1}$. We can compensate for this by considering the difference $(x+a)^{m+1} - a^{m+1}$ in which case the terms with a^{m+1}

cancel. In order to emphasize this analogy, introduce the following notation. Let a be the sequence α_1, α_2, ... and let $f(t)$ be the polynomial $a_0 + a_1 t + \cdots + a_k t^k$. Then $f(a)$ denotes the number $a_0 + a_1\alpha_1 + \cdots + a_k\alpha_k$, i.e. t^k is replaced by α_k. In particular, $a^m = \alpha_m$, since replacing t^m by α_m we obtain α_m. Analogously, $(x+a)^m = x^m + C_m^1 x^{m-1}\alpha_1 + \cdots + C_m^m \alpha^m$: we expand $(x+a)^m$ in powers of t and replace t^k by α_k. The relation (5) with this notation can be written in the form

$$g_m(x) = \frac{1}{m+1}\big((a+x)^{m+1} - a^{m+1}\big). \tag{6}$$

Remark that $a^{m+1} = \alpha_{m+1}$. Notice that we cannot establish that the polynomial given by (6) satisfies the relation (1). In fact, we have found the general form of the polynomials which satisfy the relations (2), but those relations are only *consequences* of the relation (1). Indeed, the result depends upon the sequence α_m, which can in (6) be arbitrary, whereas Theorem 7 states that the polynomial $g_m(x)$ is unique for each m. Therefore, we have not yet solved the problem.

Among the polynomials $g_m(x)$ given by (6) we have to choose those which satisfy the relation (1). Since we already know that such polynomials exist and they are unique (for each m) we only have to find the unique sequence a which defines them. This is quite simple: it is enough to put $x = 1$ into (1). Since $g_m(0) = 0$ (the constant term of g_m is 0), we get $g_m(1) = 1$. The notation of the formula (6) yields $(a+1)^{m+1} - \alpha_{m+1} = m+1$ for $m = 0, 1, 2, \ldots$ or

$$(a+1)^m - \alpha_m = m, \quad m = 1, 2, 3, \ldots$$

Definition. The numbers B_1, B_2, B_3, ... are called *Bernoulli's numbers* if the sequence B formed by them satisfies the relations

$$(B+1)^m - B_m = m \quad \text{for } m = 1, 2, 3, \ldots$$

The above relations uniquely define the sequence of Bernoulli's numbers. Indeed, expanding the above formula, by definition we get

$$1 + mB_1 + C_m^2 B_2 + \cdots + mB_{m-1} = m, \quad m = 1, 2, \ldots \tag{10}$$

(B_m cancels). From this relation for $m = 1$ we get that $B_1 = 1/2$, and every relation that follows allows us to find B_{m-1} provided that we know all B_r's with indices $r < m-1$.

Polynomials

$$B_m(x) = \frac{1}{m+1}((B+x)^{m+1} - B_{m+1})$$

where B is the sequence of Bernoulli's numbers are called *Bernoulli's polynomials*. We have proved that if the polynomial $g_m(x)$ satisfying (1) is written in the form (6), then the sequence a which corresponds to it has to coincide with the sequence B of Bernoulli's numbers. But we know, according to Theorem 7, that such a polynomial exists. Hence it must coincide with Bernoulli's polynomial $B_m(x)$, i.e.

$$B_m(x) - B_m(x-1) = x^m,$$

and so $S_m(n) = B_m(n)$. Our problem is solved.

Bernoulli's polynomials and numbers were discovered by Jacob Bernoulli (there was a large family of mathematicians of that name). His main results belong to the second half of the 17th century, but this particular discovery appeared in a book published after his death at the beginning of the 18th century. The numbers B_n were named Bernoulli's numbers by Euler (18th century) who found many other applications of those numbers.

Putting $k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13$ into (10) we obtain the following values for the numbers B_n (verify this yourself!)

$$B_1 = \frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = 0, \quad B_4 = -\frac{1}{30}, \quad B_5 = 0, \quad B_6 = \frac{1}{42},$$
$$B_7 = 0, \quad B_8 = -\frac{1}{30}, \quad B_9 = 0, \quad B_{10} = \frac{5}{66}, \quad B_{11} = 0, \quad B_{12} = -\frac{691}{2730}$$

etc. Then we easily establish:

$$S_1(n) = \frac{n(n+1)}{2}, \quad S_2(n) = \frac{n(n+1)(2n+1)}{6}, \quad S_3(n) = \frac{n^2(n+1)^2}{4},$$
$$S_4(n) = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}, \quad S_5(n) = \frac{n^2(n+1)^2(2n^2+2n-1)}{12},$$

etc.

Problems

1. Find $B_m(-1)$.

2. Prove the formula $B^m = (B-1)^m$ for $m \geqslant 2$.

3. Derive a relation, analogous to (10), which holds for Bernoulli's numbers B_m with odd indices $m \geqslant 3$. Prove that all Bernoulli's numbers B_m with odd indices, except B_1, are equal to 0.

4. Find $S_6(n)$.

5. Prove the formula for the derivative of a polynomial of a polynomial: if $f(x)$ and $g(x)$ are polynomials, then

$$f(g(x))' = f'(g(x))g'(x).$$

6. Find $(a+x)^n$ if the sequence a has the form $a_n = q^n$ for some number n.

I. R. Shafarevich,
Russian Academy of Sciences,
Moscow, Russia

THE TEACHING OF MATHEMATICS
1999, Vol. II, 2, pp. 65–80

SELECTED CHAPTERS FROM ALGEBRA

I. R. Shafarevich

Abstract. This paper is the third part of the publication "Selected chapters of algebra", the first two being published in the previous issues of the Teaching of Mathematics, Vol. I (1998), 1–22, and Vol. II, 1 (1999), 1–30.

AMS Subject Classification: 00 A 35

Key words and phrases: Set, subset, one-to-one correspondence, combinatorics.

CHAPTER III. SET

1. Sets and subsets

The notion of a set has a somewhat different meaning in mathematics than in everyday language. The ordinary word "set" usually means a large number of certain objects[1]. In mathematics a set is an arbitrary collection of objects defined by a certain property which they all have. The objects which comprise a set are called its *elements*. So, for instance, we may talk about sets of one or two elements. A set is usually denoted by a capital letter (for example, M) and its elements by small letters (for example, a, b, ... , α, β, ...). The fact that a is an element of the set M is written in the form $a \in M$ and we also say that *a belongs* to M. If M consists of elements a_1, ... , a_n, we write $M = \{a_1, \dots, a_n\}$.

A set containing a finite number of elements is called a *finite* set, while a set containing an infinite number of elements is called an *infinite* set. The number of elements of a finite set M is denoted by $n(M)$.

In this chapter we shall mainly be concerned with finite sets. The finite sets M and M' are said to be *equivalent* (equipotent) if they have the same number of elements, i.e. if $n(M) = n(M')$. We shall now describe the method which is usually used to establish the equivalence of two sets. *One-to-one correspondence* between two sets M and M' is coupling, or pairing off, their elements into pairs (a, a'), where $a \in M$, $a' \in M'$, so that each element a of M is coupled with one and only one element a' of M', and each element a' of M' is coupled with one and

This paper is an English translation of: И. Р. Шафаревич, *Избраные главы алгебры*, Математическое образование, 3, окт.–дек. 1997, Москва, стр. 2–45. In the opinion of the editors, the paper merits wider circulation and we are thankful to the author for his kind permission to let us make this version.

[1] This is not so much true for the English language as it is for Russian. The Russian word for the set *множество* has the same root as the word *много* (many).

Fig. 1

only one element a of M. If we represent the sets M and M' graphically and draw lines connecting those elements which belong to one pair, we see that each element of M is connected to one and only one element of M' and vice versa (Fig. 1).

For instance, if $n(M) = n$ and if we numerate the elements of M as follows: $M = \{a_1, \dots, a_n\}$, we have established a one-to-one correspondence between the set M and the set N of numbers 1, 2, ... , n.

If we choose two points O and E on a straight line, then to each point A which lies on that line we can correspond the real number $\dfrac{|OA|}{|OE|}$ with the + sign if A is on the same side of O as E, and with − sign in the opposite case. This establishes a one-to-one correspondence between the set of the points of the straight line and the set of all real numbers which is usually denoted by **R**. We shall consider this in more detail in one of the subsequent chapters.

If we have a one-to-one correspondence between the sets M and M', and if the elements $a \in M$ and $a' \in M'$ are coupled in the pair (a, a') we say that the element a corresponds to the element a', and that the element a' corresponds to the element a.

Two finite sets are equivalent if and only if it is possible to establish a one-to-one correspondence between them.

This statement is so ovious, that it can hardly be called a theorem. If $n(M) = n(M') = n$, we can write our sets as follows: $M = \{a_1, \dots, a_n\}$, $M' = \{a'_1, \dots, a'_n\}$, and by forming pairs (a_i, a'_i) of elements with the same index we establish a one-to-one correspondence between M and M'. Conversely, if there exists a one-to-one correspondence between M and M', and if we write M in the form $M = \{a_1, \dots, a_n\}$, then each a_i belongs to a pair with one and only one element $a' \in M'$, and we can give it the same index, i.e. we can put $a' = a_i$. By the definition of a one-to-one correspondence, we can numerate in this way all the elements of M', and we obtain that $M' = \{a'_1, \dots, a'_n\}$.

R. Dedekind (the second part of the 19th century) who did a lot to clear the role which sets have in mathematics, thought that the above statement gives, in a hidden form, the *definition* of a positive integer. According to him, it is first necessary to define the notion of one-to-one correspondence, and then a positive integer is the general property possessed by all finite sets among which it is possible to establish one-to-one correspondence. This is probably how the notion of a positive integer was formed historically (of course, without the present terminology). For example, the notion “two” was formed, as we said in Section 1 of Chapter I, by abstracting, i.e. by considering the general property shared by the sets consisting of: two eyes, two oars in a boat, two travellers walking along the road, and more generally by all the sets which can be put into a one-to-one correspondence with one of the above.

This means that the notion of a set is the most fundamental notion of mathematics, since the notion of a positive integer is founded upon the notion of a set.

In further text we shall often construct new sets, starting with two given sets.

The *product* of sets M_1 and M_2 is the set whose elements are all the pairs (a, b), where a is an arbitrary element of M_1 and b is an arbitrary element of M_2. The product of M_1 and M_2 is denoted by $M_1 \times M_2$.

For example, if $M_1 = \{1, 2\}$, $M_2 = \{3, 4\}$, then $M_1 \times M_2$ consists of the pairs $(1, 3)$, $(1, 4)$, $(2, 3)$, $(2, 4)$.

If $M_1 = M_2$ is the set $\mathbf{R}$ of all real numbers, $M_1 \times M_2$ is the set of all pairs (a, b), where a and b are real numbers. The coordinate method in the plane establishes a one-to-one correspondence between the set $M_1 \times M_2$ and the set of all points of the plane (Fig. 2).

Fig. 2

As another example, suppose that M_1 consists of the numbers $1, 2, \dots, n$, and M_2 of the numbers $1, 2, \dots, m$. Introduce two new variables x and y and correspond to a number $k \in M_1$ the monomial x^k and to the number $l \in M_2$ the monomial y^l. An element of the set $M_1 \times M_2$ has the form (k, l) and we can correspond to it the monomial $x^k y^l$. In this way we obtain a one-to-one correspondence between the set $M_1 \times M_2$ and the set of monomials of the form $x^k y^l$, where $k = 1, \dots, n$; $l = 1, \dots, m$. In other words this is the set of monomials which stand on the right-hand side of the equality

$$(1) \qquad (x + x^2 + \cdots + x^n)(y + y^2 + \cdots + y^m) = xy + x^2y + xy^2 + \cdots + x^ny^m.$$

Hence, the set of these monomials is equivalent to the set $M_1 \times M_2$.

Analogously, let M_1, M_2, $\dots$, M_r be arbitrary sets. Their *product* is the set consisting of all sequences $(a_1, \dots, a_r)$ where the i-th place is taken by an arbitrary element of the set M_i. The product of the sets M_1, $\dots$, M_r is denoted by $M_1 \times \cdots \times M_r$.

For example, if $M_1 = M_2 = M_3$ is the set of all real numbers $\mathbf{R}$, the coordinate method in the space establishes a one-to-one correspondence between the points of the space and the set $M_1 \times M_2 \times M_3$.

But in this chapter we are considered with finite sets.

Theorem 1. *If the sets $M_1, \dots, M_r$ are finite, then the set $M_1 \times \cdots \times M_r$ is also finite, and $n(M_1 \times \cdots \times M_r) = n(M_1) \cdots n(M_r)$.*

We shall first prove the theorem for the case of two sets, i.e. when $r = 2$; this will be the induction basis. If $M_1 = \{a_1, \dots, a_n\}$, $M_2 = \{b_1, \dots, b_m\}$, then all the pairs (a_i, b_j) can be written in the form of a rectangle

$$\begin{matrix} (a_1, b_1) & \dots & (a_n, b_1) \\ (a_1, b_2) & \dots & (a_n, b_2) \\ \dots & \dots & \dots \\ (a_1, b_m) & \dots & (a_n, b_m) \end{matrix} \tag{2}$$

The j-th row above contains pairs whose last element is always b_j. In each row the number of pairs is equal to the number of all a_i's, i.e. it is n. The number of the rows is equal to the number of all b_j's, i.e. it is equal to m. Hence, the number of pairs is nm. Notice that the rectangle (2) resembles, in a way, Fig. 2. (A different line of reasoning would be to say that the set M_1 is equivalent to the set $\{1, \dots, n\}$ or to the set of monomials $\{x, x^2, \dots, x^n\}$ and that M_2 is equivalent to the set $\{y, y^2, \dots, y^m\}$. Then, $n(M_1 \times M_2)$ is, as we have seen, the number of terms in the right-hand side of (1). Putting $x = 1$, $y = 1$, we conclude that this number of terms is nm.)

The proof of the general case of r sets $M_1, \dots, M_r$ will be carried out by induction on r. In each sequence $(a_1, \dots, a_r)$ we introduce two more brackets, and write it in the form $((a_1, \dots, a_{r-1}), a_r)$. Clearly, this does not alter the number of the sequences. But the sequence $((a_1, \dots, a_{r-1}), a_r)$ is the pair (x, a_r), where $x = (a_1, \dots, a_{r-1})$ can be considered to be an element of the set $M_1 \times \cdots \times M_{r-1}$. Hence, the set $M_1 \times \cdots \times M_r$ is equivalent to the set $P \times M_r$, where $P = M_1 \times \cdots \times M_{r-1}$. We have proved that $n(P \times M_r) = n(P)n(M_r)$, and by the induction hypothesis we have $n(P) = n(M_1) \cdots n(M_{r-1})$. Therefore, $n(M_1 \times \cdots \times M_r) = n(M_1) \cdots n(M_{r-1})n(M_r)$ and the proof is complete.

Using Theorem 1 we can once more form the expression for the number of divisors of a positive integer n. Suppose that n has the cannonical representation

$$n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}.$$

In Section 3 of Chapter I we saw that the divisors of n can be written in the form

$$m = p_1^{\beta_1} \cdots p_r^{\beta_r},$$

where β_i can take any integral value between 0 and α_i (formula (11) of Chapter I). In other words, the set of divisors is equivalent to the set of sequences $(\beta_1, \dots, \beta_r)$ where β_i takes the above mentioned values. But this is exactly the product $M_1 \times \cdots \times M_r$ of the sets M_i where M_i is the set $\{0, 1, \dots, \alpha_i\}$. Since $n(M_i) = \alpha_i + 1$, according to Theorem 1 the number of divisors is $(\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_r + 1)$. This formula was derived in a different way in Section 3, Chapter I.

If the sets $M_1, \ldots, M_r$ coincide, i.e. if $M_1 = M_2 = \cdots = M_r = M$ their product $M_1 \times \cdots \times M_r$ is denoted by M^r. Consider the case when $M_1 = \cdots = M_r = I$, and the set I has two elements a and b. An element of I^r is a sequence of r symbols, each one being a or b, e.g. $aababbba$ (for shortness sake we omit the commas). This can be considered as a word of r letters written in the alphabet of two letters, a being a dot and b a dash. Therefore, $n(I^r)$ is equal to the number of words of length r, written in Morse's alphabet. As we see, it is equal to 2^r (all $n_i = 2$).

In further text we consider sets contained in a given set M. They are called its *subsets*. This means that a subset N of a set M contains only elements of M, but not necessarily all of them. The fact that N is a subset of M is written as $N \subset M$. We also take that M is a subset of itself. As we shall see later, it is very convenient to consider the subset of M containing no elements—this simplifies greatly many definitions and theorems. This subset is callled the *empty subset* and is denoted by $\varnothing$. By definition we take $n(\varnothing) = 0$.

If $N \subset M$, the set of all elements of M which do not belong to N is called the *complement* of N and is denoted by $\overline{N}$. For instance, if M is the set of all positive integers, and if N is the set of all even positive integers, then $\overline{N}$ is the set of all odd positive integers. If $N = M$, then $\overline{N} = \varnothing$.

If N_1 and N_2 are two subsets of M (i.e. $N_1 \subset M$ and $N_2 \subset M$) then the set of all elements which belong to N_1 and N_2 is called their *intersection* and is denoted by $N_1 \cap N_2$. For example, if M is the set of all positive integers, if N_1 is the subset of all those divisible by 2, and N_2 the subset of all those divisible by 3, then $N_1 \cap N_2$ is the set of all positive integers divisible by 6.

If N_1 and N_2 do not have common elements, then by definition $N_1 \cap N_2 = \varnothing$, the empty set. So, if M and N_1 are the same as in the previous example, and N_2 is the set of odd positive integers, then $N_1 \cap N_2 = \varnothing$.

The set containing elements which belong to the subset N_1 or the subset N_2 is called their *union* and is denoted by $N_1 \cup N_2$. For example, if M is again the set of all positive integers, and N_1 and N_2 are the subsets of all even and odd numbers, respectively, then $N_1 \cup N_2 = M$.

a) b)

Fig. 3

Intersections and unions of sets can be represented graphically as in Figure 3. In Fig. 3a) $M_1 \cup M_2$ is hatched by horizontal and $M_1 \cap M_2$ by vertical lines. In Fig. 3b) the set $(M_1 \cup M_2) \cap M_3$ is hatched.

In this chapter we shall consider subsets of a finite set M, which satisfy certain conditions and we shall derive formulas for the number of all such subsets. The branch of mathematics concerned with such questions is called *combinatorics.*

Therefore, combinatorics is the theory of arbitrary finite sets. We do not use notions such as distance or the magnitude of an angle, equation or its roots, but only the notion of a subset and the number of its elements. Hence, it is very surprising that, using only such miserly material, we can find many regularities and connections with other branches of mathematics which are not at all obvious.

Problems

1. Let $M = M'$ be the set of all positive integers. Couple into pairs the number $a \in M$ with $b \in M'$ such that $b = 2a$. Is this a one-to-one correspondence between M and M'?

2. Let N be the set of all positive integers, let $M = N \times N$ and let M' be the set of positive rational numbers. Couple into pairs $(n_1, n_2) \in M$ with $a \in M'$ if $a = n_1/n_2$. Is this a one-to-one correspondence?

3. How many different one-to-one correspondences exist between two sets M and M' if $n(M) = n(M') = 3$? Draw them analogously as in Fig. 1.

4. Every one-to-one correspondence between the sets M and M' defines the set of those pairs (a, a'), where $a \in M$ and $a' \in M'$ correspond to each other, i.e. it defines a subset $\Gamma \subset M \times M'$ which is called the *graph of correspondence.* Let Γ_1 and Γ_2 be graphs of two one-to-one correspondences. Prove that $\Gamma_1 \cap \Gamma_2$ is a graph of a one-to-one correspondence if and only if $\Gamma_1 = \Gamma_2$ and the two given correspondences coincide.

5. Let $n(M) = n(M') = n$ and let Γ be the graph of a one-to-one correspondence between M and M' (see Problem 4). Evaluate $n(\Gamma)$.

6. Let M be the set of all positive integers, let $N_1 \subset M$ be the subset of all numbers divisible by a given number a_1 and let $N_2 \subset M$ be the subset of all numbers divisible by a given number a_2. Describe the sets $N_1 \cup N_2$ and $N_1 \cap N_2$.

7. Prove that $\overline{(\overline{N})} = N$, i.e. that the complement of the complement of a subset N is exactly N.

2. Combinatorics

We start with the simplest question: determine the number of all subsets of a finite set.

We first solve the problem for small values of $n(M)$. We will write down the subsets N of M writing in one row all the subsets with the same number of elements (i.e. with the same value of $n(N)$). The rows are arranged in the ascending order of $n(N)$.

1.	$n(M) = 1,$	$M = \{a\}$		
	$n(N) = 0;$	$N = \varnothing$		
	$n(N) = 1;$	$N = M = \{a\}$		
2.	$n(M) = 2,$	$M = \{a, b\}$		
	$n(N) = 0;$	$N = \varnothing$		
	$n(N) = 1;$	$N = \{a\},$	$N = \{b\}$	
	$n(N) = 2;$	$N = M = \{a, b\}$		
3.	$n(M) = 3,$	$M = \{a, b, c\}$		
	$n(N) = 0;$	$N = \varnothing$		
	$n(N) = 1;$	$N = \{a\},$	$N = \{b\},$	$N = \{c\}$
	$n(N) = 2;$	$N = \{a, b\},$	$N = \{a, c\},$	$N = \{b, c\}$
	$n(N) = 3;$	$N = M = \{a, b, c\}$		

Table 1

We see that if $n(M) = 1$, the number of subsets is 2, if $n(M) = 2$ it is 4 and if $n(M) = 3$ it is 8. This suggests the general statement.

Theorem 2. *The number of all subsets of a finite set M is $2^{n(M)}$.*

There is a general method which reduces the investigation of an arbitrary finite set to the investigation of sets with smaller number of elements. The set M is called the *sum* of its two subsets $M_1 \subset M$ and $M_2 \subset M$ if $M_1 \cup M_2 = M$, $M_1 \cap M_2 = \varnothing$. Clearly, this is equivalent to $M_2 = \overline{M_1}$ and $M_1 = \overline{M_2}$. In this case each element of M belongs to one of the subsets M_1 or M_2 (since $M_1 \cup M_2 = M$) and only to one of them (since $M_1 \cap M_2 = \varnothing$). Hence, $n(M) = n(M_1) + n(M_2)$. The fact that M is the sum of M_1 and M_2 is written as $M = M_1 + M_2$. Such a representation is also called a *partition* of M into M_1 and M_2.

Let $M = M_1 + M_2$ and let $N \subset M$ be an arbitrary subset. Then any element $a \in N$ belongs either to M_1 (in this case $a \in N \cap M_1$) or to M_2 (in this case $a \in N \cap M_2$), and only one of these cases can take place (since $M_1 \cap M_2 = \varnothing$). Hence $N = (N \cap M_1) + (N \cap M_2)$. Conversely, if $N_1 \subset M_1$ and $N_2 \subset M_2$ are arbitrary subsets, then $N_1 \subset M$, $N_2 \subset M$ and $N = N_1 \cup N_2 \subset M$, whereas $N \cap M_1 = N_1$ and $N \cap M_2 = N_2$. In this way we establish a one-to-one correspondence between the subsets N of M and the pairs (N_1, N_2) where N_1 and N_2 are arbitrary subsets of M_1 and M_2, respectively.

We now formulate this result in terms of sets. Denote by $U(M)$ the set of all subsets of a set M. One should not be alarmed because we consider here subsets as elements of a new set. So, for example, associations of civil or electrical engineers are elements of the general association of engineers. In Table 1 we decribed the sets $U(M)$ when $n(M) = 1$, 2 or 3. The result obtained above can be formulated as follows: if $M = M_1 + M_2$ is a partition of M, then the set $U(M)$ is in a one-to-one correspondence with the set $U(M_1) \times U(M_2)$. Denote the number $n(U(M))$ by $v(M)$—this is the required number of all subsets. Applying Theorem 1 we deduce that

$$v(M_1 + M_2) = v(M_1)v(M_2). \tag{3}$$

The equality (3) reduces the evaluation of $v(M)$ to the evaluation of $v(M_1)$ and $v(M_2)$ for the sets M_1 and M_2 with smaller number of elements. In order to obtain the final result, consider the partition of M not into two, but into an arbitrary number of subsets. We can define this concept inductively, saying that $M = M_1 + \cdots + M_r$ if $M = (M_1 + \cdots + M_{r-1}) + M_r$, where the expression $M_1 + \cdots + M_{r-1}$ is taken to be already defined. In fact, when we say that $M = M_1 + \cdots + M_r$, this means that $M_1, \ldots, M_r$ are subsets of M and that every element of M belongs to one and only one of the subsets $M_1, \ldots, M_r$. For example, if M is the set of all positive integers, then $M = M_1 + M_2 + M_3$, where M_1 is the subset of all numbers divisible by 3, M_2 is the subset of all numbers of the form $3r + 1$ and M_3 is the subset of all numbers of the form $3r + 2$.

From (3), for finite sets M_i we obtain, by induction

$$v(M_1 + \cdots + M_r) = v(M_1) \cdots v(M_r). \tag{4}$$

If $n(M) = n$, then there exists the "tiniest" partition of M into n subsets M_i, each having only one element, i.e. $M = M_1 + \cdots + M_n$. If $M = \{a_1, \ldots, a_n\}$, then $M_i = \{a_i\}$. The one element set M_i has two subsets: the empty set $\varnothing$ and M_i itself (see Table 1, first row). Hence, $v(M_i) = 2$ and applying formula (4) to the partition $M = M_1 + \cdots + M_n$ we obtain that $v(M) = 2^n$, as stated in Theorem 2.

The question of the number of all subsets of a given set appears in connection with certain problems regarding numbers. For example, consider the following question: in how many ways can a positive integer n be written as a product of two relatively prime factors? Let $n = ab$, where a and b are relatively prime and let $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$ be the canonical prime factorization. Then a and b are divisors of n and, as we saw in Section 3 of Chapter I, each one of them has the form $p_1^{\beta_1} \cdots p_r^{\beta_r}$ where $0 \leqslant \beta_i \leqslant \alpha_i$. But since a and b are relatively prime, then if some p_i divides a, then it cannot divide b and hence appears in a with degree α_i. Therefore, in order to obtain the required factorization $n = ab$, it is necessary to choose an arbitrary subset N of the set $M = \{p_1, \ldots, p_r\}$ and to equate a to the product of $p_i^{\alpha_i}$ for $p_i \in N$. Then a divides n and $n = ab$ is the required factorization. According to Theorem 2, the number of all factorizations of n into products of two relatively primer factors is 2^r, where r is the number of different prime factors of n.

It should be noted that in the above evaluation we considered the factorization $n = ab$ and $n = ba$ to be different. In fact, if a, and hence the factorization $n = ab$, corresponds to the subset $N \subset \{p_1, \ldots, p_r\}$, then b corresponds to the subset consisting of those $p_i \in M$ which do not belong to N, i.e. which belong to the complement $\overline{N}$ of N. Therefore, in our evaluation we corresponded the factorizations $n = ab$ and $n = ba$ to two different subsets N and $\overline{N}$. Hence, if we do not want to make a difference between the factorizations $n = ab$ and $n = ba$, then the two subsets N and $\overline{N}$ should be treated as one, and then the number of all factorizations in this sense would be 2^{r-1}.

We now pass on to a more subtle problem: find the number of subsets of a given finite set M which contain m elements and m is a given number. In order to do this, we again collect all the subsets $N \subset M$ such that $n(N) = m$ into one set denoted by $U(M,m)$. If we put $n(U(M,m)) = v(M,m)$, then this is the number which we wish to find. In Table 1 we wrote the sets which belong to $U(M,m)$ on one row. Hence, we obtain the values of $v(M,m)$ for small values of $n(M)$:

$$\begin{array}{lllll}
n(M) = 1: & v(M,0) = 1, & v(M,1) = 1 & & \\
n(M) = 2: & v(M,0) = 1, & v(M,1) = 2, & v(M,2) = 1 & \\
n(M) = 3: & v(M,0) = 1, & v(M,1) = 3, & v(M,2) = 3, & v(M,3) = 1
\end{array}$$

Table 2

Theorem 3. *If $n(M) = n$, the number of subsets $N \subset M$ of the set M which contain m elements (i.e. such that $n(N) = m$) is equal to the binomial coefficient C_n^m. In other words, $v(M,m) = C_n^m$.*

The proof is based upon the same idea as the proof of Theorem 2. Namely, suppose that the set M is the sum of two subsets: $M = M_1 + M_2$ and we shall express the number $v(M,m)$ in terms of the numbers $v(M_1,m)$ and $v(M_2,m)$. If $M = M_1 + M_2$, then each subset $N \subset M$ can be written in the form $N = N_1 + N_2$, where $N_1 = N \cap M_1$, $N_2 = N \cap M_2$. If we take into account the condition $n(N) = m$, then we must have $n(N_1) + n(N_2) = m$. Let k and l be two nonnegative integers such that $k + l = m$. Consider all subsets $N \subset M$ such that $n(N \cap M_1) = k$, and $n(N \cap M_2) = l$, denote the set of all these subsets by $U(k,l)$ and put $n(U(k,l)) = v(M,k,l)$. Then in the same way as in the proof of Theorem 2 we see that

$$v(M,k,l) = v(M_1,k)v(M_2,l). \tag{5}$$

The set $U(M,m)$ can clearly be partitioned into sets $U(M,k,l)$ for various pairs of numbers k, l, such that $k + l = m$. Therefore the number of its elements $v(M,m)$ is equal to the sum of all numbers $v(M,k,l)$ for all k and l such that $k + l = m$, i.e. for all the values: $k = 0$, $l = m$; $k = 1$, $l = m - 1$; $\ldots$; $k = m$, $l = 0$. From the relation (5) we obtain

$$v(M,m) = v(M_1,m)v(M_2,0) + v(M_1,m-1)v(M_2,1) + \cdots + v(M_1,0)v(M_2,m). \tag{6}$$

Of course, if in the product $v(M_1,k)v(M_2,l)$ it turns out that $k > n(M_1)$, we have to take $v(M_1,k) = 0$ and the same holds for $v(M_2,l)$.

We have obtained a relation analogous to the relation (3), although it is more complicated.

We have met the relation (6) in connection with a completely different problem. This is, in fact, the coefficient of x^m in the product of two polynomials $f(x)$ and $g(x)$ if the coefficient of x^k in $f(x)$ is $v(M_1,k)$ and the coefficient of x^l in $g(x)$

is $v(M_2, l)$; see formula (1) od Chapter II. In order to establish the connection between these two statements, define for an arbitrary finite set M the polynomial $f_M(x)$ whose coefficients are $v(M, s)$:

$$f_M(x) = v(M, 0) + v(M, 1)x + \cdots + v(M, n)x^n, \tag{7}$$

where $n = n(M)$.

For instance, according to Table 2, if $n(M) = 1$, then $f_M(x) = 1 + x$, if $n(M) = 2$ then $f_M(x) = 1 + 2x + x^2$, if $n(M) = 3$, then $f_M(x) = 1 + 3x + 3x^2 + x^3$. Now comparing the relations (6) and (7) we can write

$$f_M(x) = f_{M_1}(x) \cdot f_{M_2}(x), \quad \text{if } M = M_1 + M_2. \tag{8}$$

Hence, if we introduce polynomials $f_M(x)$ instead of the numbers $v(M)$ we obtain a complete similarity with the formula (3). We see that the polynomial $f_M(x)$ turns out to be a just replacement for the number $v(M)$ in our more complicated problem. This is not a rare thing to happen. If we have to deal not with one number, but with a finite sequence of numbers $(a_0, \ldots, a_n)$, then its properties are often well expressed by means of the polynomial $a_0 + a_1 x + \cdots + a_n x^n$. We shall see that later, in other examples.

It remains literally to repeat the end of the proof of Theorem 2. If $M = M_1 + \cdots + M_r$, then from (8) we obtain, by induction,

$$f_M(x) = f_{M_1}(x) \cdots f_{M_r}(x).$$

Now put $n(M) = n$ and partition the set M into n subsets each containing one element: $M = M_1 + \cdots + M_n$, $n(M_i) = 1$. The one element set M_i has two subsets: the empty set $\varnothing$ with $n(\varnothing) = 0$ and M_i with $n(M_i) = 1$. Therefore, $v(M_i, 0) = 1$, $v(M_i, 1) = 1$, $v(M_i, k) = 0$ for $k > 1$, $f_{M_i} = 1 + x$ and we conclude that for any finite set M we have

$$f_M(x) = (1 + x)^{n(M)}.$$

The expression $(1 + x)^{n(M)}$ can be written in the form of a polynomial in x by means of the binomial formula. We have seen (formulas (20) and (24) of Chapter II) that for $n = n(M)$:

$$(1 + x)^n = C_n^0 + C_n^1 x + C_n^2 x^2 + \cdots + C_n^n x^n, \quad \text{where } C_n^m = \frac{n!}{m!\,(n - m)!}.$$

Therefore, recalling the definition of the polynomial $f_M(x)$ (formula (7)), we obtain

$$v(M, m) = C_n^m = \frac{n!}{m!\,(n - m)!} \quad \text{for } n = n(M). \tag{9}$$

This is the answer to our question.

By counting the subsets of M containing 0, 1, 2, ... , n elements where $n = n(M)$, we have counted all the subsets of M. Therefore, $v(M, 0) + v(M, 1) + \cdots + v(M, n) = v(M)$, or using (9) and Theorem 2, $C_n^0 + C_n^1 + \cdots + C_n^n = 2^n$. This relation for the binomial coefficients is easily obtained from the binomial formula, as we have done in Section 3 of Chapter II.

A subset of m elements of the set $\{a_1, \dots, a_n\}$ is sometimes called a *combination* of n elements, taken m at a time. Hence, the binomial coefficient C_n^m is the number of all such combinations.

The above question of the number of subsets $N \subset M$, if $n(M) = n$, $n(N) = m$, is connected with some questions regarding positive integers. For example, consider the question: in how many ways can we write a positive integer n in the form of r summands, where r is a given number? In other words, what is the number of solutions of the equation $x_1 + \cdots + x_r = n$ in positive integers $x_1, \dots, x_r$? Solutions with different order of the unknowns are considered to be different. For example, if $n = 4$, $r = 2$, we have $4 = 1 + 3 = 2 + 2 = 3 + 1$, and hence there are three solutions: $(1, 3)$, $(2, 2)$, $(3, 1)$.

Consider the segment AB of length n. Its points whose distance from the initial point A are integers will be called integral. Clearly, to each solution of the equation $x_1 + \cdots + x_r = n$ corresponds a partition of the segment AB into r segments with integral end points of length x_1, x_2, ... , x_r (Fig. 4).

Fig. 4

In its turn, such a partition is defined by the end points of the first $r - 1$ segments (the end point of the last one is B). These end points define a subset N of the set M of integral points of the segment AB which are different from B. Clearly, $n(N) = r - 1$, and in this way we have defined a one-to-one correspondence between the integer solutions of the equation $x_1 + \cdots + x_r = n$ and the subsets $N \subset M$, where $n(N) = r - 1$, $n(M) = n - 1$. Therefore, the number of such solutions is equal to the number of such subsets. Applying formula (9) we conclude that the number of these subsets is C_{n-1}^{r-1}. If we do not fix the number of summands into which the number n is decomposed, then the number of all partitions is evidently equal to the sum of partitions into r summands for $r = 1, 2, \dots, n$. Therefore, the number of partitions is equal to the sum of all binomial coefficients C_{n-1}^{r-1} where $r = 1, 2, \dots, n$. We know that this sum is 2^{n-1}. In other words, a positive integer n can be partitioned into integer summands in 2^{n-1} ways (if we allow arbitrary number of summands, and take into account their order).

Return now to the derivation of formula (9). The method used—the introduction of the polynomials $f_M(x)$—turns out to be very useful in other cases, and we shall come back to it later. But formula (9) which connects the numbers $v(M, m)$ with binomial coefficients can be derived in another way. Consider the expression $(1 + x)^n$ as the product of n equal factors

$$(1 + x)^n = (1 + x)(1 + x) \cdots (1 + x) \tag{10}$$

and let us expand the product on the right-hand side of (10). We numerate its factors, i.e. we give them numbers 1, 2, ... , n which form the set $M = \{1, 2, \dots, n\}$.

In order to expand the product (10) we have to multiply each time n terms 1 or x, taking them from one of the brackets. Hence, each term of the expanded expression (10) is defined by indicating from which brackets with m numbers $i_1, i_2, \dots, i_m$. Then 1 is taken from the remaining $n-m$ brackets, and as a result we obtain the term x^m. We see that each term of the expanded expression (10) is defined by the subset $N = \{i_1, \dots, i_m\}$ of M which gives the number of brackets from which x is taken. From the remaining brackets we take 1. The remaining brackets have those numbers which belong to the complement $\overline{N}$ of N. Therefore, the number of appearences of the term x^m is equal to the number of subsets $N \subset M$ containing m elements, and this is $v(M,m)$. Hence, the expression (10) in the expanded form is the sum of the terms of the form $v(M,m)x^m$:

$$(1+x)^n = v(M,0) + v(M,1)x + \cdots + v(M,n)x^n.$$

Comparing this with the definition of binomial coefficients (formula (20) of Chapter II) we obtain a new proof of the equality $v(M,m) = C_n^m$.

The same reasoning can be applied to a more general case. Consider the product of first degree polynomials $x + a_i$, where the coefficient of x is 1. Let us try to write the product

$$(x+a_1)(x+a_2)\cdots(x+a_n) \tag{11}$$

in the form of a polynomial in x. As before we numerate the n factors. Then each term in the expanded product (11) is obtained by taking $a_{i_1}, a_{i_2}, \dots, a_{i_m}$ from the factors numerated $i_1, i_2, \dots, i_m$ and taking x from the remaining $n-m$ factors. The obtained term has the form $a_{i_1}a_{i_2}\cdots a_{i_m}x^{n-m}$ and if all the terms of degree $n-m$ are collected together we get $\sigma_m(a_1, \dots, a_n)x^{n-m}$ where $\sigma_m(a_1, \dots, a_n)$ is the sum of all products of the form $a_{i_1}\cdots a_{i_m}$ where $\{i_1, \dots, i_m\}$ runs over all sets of indices formed from $1, \dots, n$. Hence, the polynomial $\sigma_m(a_1, \dots, a_n)$ has C_n^m terms. For example, $\sigma_1(a_1, \dots, a_n) = a_1 + \cdots + a_n$, and $\sigma_2(a_1, \dots, a_n) = a_1a_2 + a_1a_3 + \cdots + a_2a_3 + \dots a_{n-1}a_n$—it is the sum of all products a_ia_j with $i < j$. The last polynomial σ_n has the form $\sigma_n(a_1, \dots, a_n) = a_1\cdots a_n$. This is the first time that we encounter polynomials in an arbitrary number n of variables. Polynomials $\sigma_1, \dots, \sigma_n$ have a very important role in algebra. In particular, we have proved the formula

$$\begin{multline} (x+a_1)\cdots(x+a_n) = \\ x^m + \sigma_1(a_1, \dots, a_n)x^{n-1} + \sigma_2(a_1, \dots, a_n)x^{n-2} + \cdots + \sigma_n(a_1, \dots, a_n). \end{multline} \tag{12}$$

It is called Viète's formula.

Viète's formula expresses an important property of polynomials. Suppose that the polynomial $f(x)$ of degree n has n roots $\alpha_1, \dots, \alpha_n$. Then, as we have seen more than once, it is divisible by the product $(x-\alpha_1)\cdots(x-\alpha_n)$, and since this product is also of degree n, then $f(x) = c(x-\alpha_1)\cdots(x-\alpha_n)$, where c is a number. Suppose that the coefficient of the leading term of $f(x)$ is 1. Then the number c must also be 1, and we have

$$f(x) = (x-\alpha_1)\cdots(x-\alpha_n).$$

We can apply Viète's formula (12) to it by putting $a_i = -\alpha_i$. Since all the terms of the polynomial σ_k are products of k variables taken from $a_1, \dots, a_n$, then replacing a_i by $-\alpha_i$ gives rise to a factor $(-1)^k$, namely: $\sigma_k(-\alpha_1, \dots, -\alpha_n) = (-1)^k \sigma_k(\alpha_1, \dots, a_n)$. Hence, from (12) we obtain

$$(13)\quad (x-\alpha_1)\cdots(x-\alpha_n) = x^n - \sigma_1(\alpha_1,\dots,\alpha_n)x^{n-1} + \cdots + (-1)^n\sigma_n(\alpha_1,\dots,\alpha_n).$$

This formula expresses the coefficients of the polynomial $f(x) = (x-\alpha_1)\cdots(x-\alpha_n)$ in terms of its roots and it is also called Viète's formula. You know its special case for the quadratic equation: in that case there are only two polynomials σ_1 and σ_2, $\sigma_1 = \alpha_1 + \alpha_2$, $\sigma_2 = \alpha_1\alpha_2$.

In conclusion, consider again formula (9) for the number of subsets (or the number of combinations). We deduced it from the binomial formula, which was, in turn, proved in Section 3 of Chapter II using the properties of the derivative. That is a rather involved method. It would be desirable to have another proof of this formula based only upon combinatorial reasoning. We shall give such a proof of an even more general formula. Notice that each subset N of the set M defines a partition $M = N + \overline{N}$ where $\overline{N}$ is the complement of N. We consider a more general case: an arbitrary partition $M = M_1 + \cdots + M_r$ into subsets with prescribed number of elements: $n(M_1) = n_1, \dots, n(M_r) = n_r$. The sequence $(n_1, \dots, n_r)$ will be called the *type* of the partition $M = M_1 + \cdots + M_r$. We suppose that none of the sets M_i is empty, i.e. that all $n_i > 0$.

Since we are dealing all the time with one and only set M where $n(M) = n$, it shall not always be present in our notations. Denote the number of all possible partitions of our set M which have the prescribed type $(n_1, \dots, n_r)$ by $C(n_1, \dots, n_r)$. Of course, we must have $n_1 + \cdots + n_r = n$. Notice also that we are taking into account the order of the sets $M_1, \dots, M_r$. For instance, for $r = 2$ and given n_1 and n_2, $n_1 + n_2 = n$, we take the partitions $M = M_1 + M_2$ and $M = M_2 + M_1$, with $n(M_1) = n_1$ and $n(M_2) = n_2$, to be different. Indeed, if $n_1 \neq n_2$ these partitions are of different types. Owing to this, each partition $M = M_1 + M_2$ defines one subset M_1 (the first one) and we have a connection with the previously considered problem: $C(n_1, n_2) = v(M, n_1)$. In other words, for any $m < n$, we have $v(M, m) = C(m, n-m)$. We shall now derive an explicit formula for the number $C(n_1, \dots, n_r)$. Consider an arbitrary partition $M = M_1 + \cdots + M_r$ of type $(n_1, \dots, n_r)$. Suppose that at least one of the numbers $n_1, \dots, n_r$ is different from 1. For instance, suppose that $n_1 > 1$ and choose an arbitrary element $a \in M_1$. Denote by M_1' the set of all elements of M_1 different from a (this is the complement of the set $\{a\}$ taken as a subset of M_1). Then we have the partition $M_1 = M_1' + \{a\}$ and to our partition $M = M_1 + \cdots + M_r$ corresponds a new partition $M = M_1' + \{a\} + M_2 + \cdots + M_r$ of type $(n_1 - 1, 1, n_2, \dots, n_r)$. In this way from all partitions of type $(n_1, n_2, \dots, n_r)$ we obtain all partitions of type $(n_1 - 1, 1, n_2, \dots, n_r)$: the partition $M = M_1' + \{a\} + M_2 + \cdots + M_r$ is obtained from the partition $M = M_1 + \cdots + M_r$, where $M_1 = M_1' + \{a\}$. Moreover, one partition of type $(n_1, n_2, \dots, n_r)$ gives rise to n_1 different partitions of type $(n_1 - 1, 1, n_2, \dots, n_r)$, depending on the choice of $a \in M_1$. Hence, we have

$$(14)\qquad n_1 C(n_1, n_2, \dots, n_r) = C(n_1 - 1, 1, n_2, \dots, n_r).$$

Applying the same method to partitions of type $(n_1 - 1, 1, n_2, \ldots, n_r)$ we obtain that $(n_1 - 1)C(n_1 - 1, 1, n_2, \ldots, n_r) = C(n_1 - 2, 1, 1, n_2, \ldots, n_r)$, i.e. that $n_1(n_1 - 1)C(n_1, n_2, \ldots, n_r) = C(n_1 - 1, 1, 1, n_2, \ldots, n_r)$ and

$$n_1!\, C(n_1, n_2, \ldots, n_r) = C(\underbrace{1, \ldots, 1}_{n_1 \text{ times}}, n_2, \ldots, n_r).$$

We now apply the same reasoning to the parameter n_2 in $C(1, \ldots, 1, n_2, \ldots, n_r)$. In the same way as before we obtain the relation $n_2!\, C(1, \ldots, 1, n_2, n_3, \ldots, n_r) = C(1, \ldots, 1, n_3, \ldots, n_r)$ where 1 appears in the first $n_1 + n_2$ places, that is to say

$$n_1!\, n_2!\, C(n_1, n_2, \ldots, n_r) = C(\underbrace{1, \ldots, 1}_{n_1 + n_2 \text{ times}}, n_3, \ldots, n_r).$$

Finally, if we apply the procedure to all the parameters n_1, n_2, ... , n_r we obtain the formula

$$n_1!\, n_2! \cdots n_r!\, C(n_1, n_2, \ldots, n_r) = C(\underbrace{1, 1, \ldots, 1}_{n \text{ times}}), \tag{15}$$

since $n_1 + n_2 + \cdots + n_r = n$. It remains to find the value of the expression $C(1, \ldots, 1)$. In order to do so notice that the above formula was proved for partitions of all types $(n_1, \ldots, n_r)$. Apply it to the simplest type (n). There is only one partition of this type, namely $M = M$, and so $C(n) = 1$. On the other hand, formula (15) gives

$$n!\, C(n) = C(\underbrace{1, 1, \ldots, 1}_{n \text{ times}}).$$

Therefore $C(1, \ldots, 1) = n!$ and substituting this into (15) we obtain the final expression

$$C(n_1, \ldots, n_r) = \frac{n!}{n_1!\, n_2! \cdots n_r!}, \quad \text{where } n = n_1 + \cdots + n_r. \tag{16}$$

For $n = 2$ instead of (n_1, n_2), $n_1 + n_2 = n$ it is more usual to write $(n, m - n)$. Since $C(m, n - m) = v(M, m)$, formula (16) reduces to the relation (9).

Remark 1. Consider again the expression $C(1, \ldots, 1)$ which appeared at the end of the above proof. What is a partition of type $(1, \ldots, 1)$? It is a partition of M into one element sets. But recall that we must take into account the order of the sets in the partition $M = M_1 + \cdots + M_r$. Hence, a partition $M = \{a_1\} + \cdots + \{a_n\}$ gives a numertaion of the elements of M. The number $C(1, \ldots, 1)$ shows in how many ways we can numerate the elements of M. It can be said that $C(1, \ldots, 1)$ gives the number of different arrangements of the elements of M. As we know, the number of such arrangements is $n!$. Various arrangements are also called *permutations*. For example, if $M = \{a, b, c\}$, which means that $n = 3$, we have 6 permutations

$$(a, b, c), \quad (a, c, b), \quad (b, a, c), \quad (b, c, a), \quad (c, a, b), \quad (c, b, a).$$

Remark 2. In the case $r = 2$, the expression $C(n_1, n_2)$ coincides with the binomial coeffcient—we have already given two proofs of this fact. An analogous interpretation has the expression $C(n_1, \dots, n_r)$ for any r. It can be proved that if $x_1, \dots, x_r$ are variables, then in the expansion of $(x_1 + \dots + x_r)^n$ we obtain terms of the form $x_1^{n_1} \cdots x_r^{n_r}$ with $n_1 + \dots + n_r = n$, n_i nonnegative integers, and that the coefficient of $x_1^{n_1} \cdots x_r^{n_r}$ is $C(n_1, \dots, n_r)$. We have only to return to our first definition of a partition, allowing the empty set to appear among M_i's and hence allowing zero to be among the numbers n_i. It is easily seen that (16) remains valid in this case also, provided we take $0! = 1$. The proof of this generalization of the binomial formula to the case of r variables is perfectly analogous to the second (combinatorial) proof of the relation $v(M, m) = C_n^m$ (where $n = n(M)$) given above.

For instance, this formula gives that $(x_1 + x_2 + x_3)^3$ is equal to the sum of terms $C(n_1, n_2, n_3) x_1^{n_1} x_2^{n_2} x_3^{n_3}$ where (n_1, n_2, n_3) runs over all triplets of nonnegative integers such that $n_1 + n_2 + n_3 = 3$, and $C(n_1, n_2, n_3)$ is evaluated by formula (16) (with the condition $0! = 1$). Substitution gives

$$(x_1 + x_2 + x_3)^3 = \\ x_1^3 + x_2^3 + x_3^3 + 3x_1^2 x_2 + 3x_1 x_2^2 + 3x_1^2 x_3 + 3x_1 x_3^2 + 3x_2^2 x_3 + 3x_2 x_3^2 + 6x_1 x_2 x_3.$$

Problems

1. Let $I = \{p, q\}$ be a set containing two elements and let $M = \{a_1, \dots, a_n\}$ be a set of n elements. To each subset $N \subset M$ correspond the following element: a word from I^n where on the i-th place stands p if $a_i \in N$, and q if a_i does not belong to N. Prove that this establishes a one-to-one correspondence between the sets $U(M)$ and I^n. Use this to derive Theorem 2 from Theorem 1.

2. How can the intersection and the union of subsets N_1 and N_2 of the set M be expressed in terms of their corresponding words from I^n (see Problem 1)?

3. Find the number of all partitions $M_1 + \dots + M_r$ of all types, but for a fixed number r. Verify that for $r = 2$ the answer is given by Theorem 2.

4. Find the sum of all numbers $C(n_1, \dots, n_r)$ for all $n_i \geqslant 0$, $n_1 + \dots + n_r = n$, for given r and n. Give two solutions: one based upon Problem 3 and the other based upon the statement given in Remark 2.

5. Find the number of factorizations of a given positive integer n into a product of r factors: $n = a_1 \cdots a_r$, which are mutually relatively prime.

6. What is the number of solutions of the equation $x_1 + \dots + x_r = n$, for given n and r, in integers $x_i \geqslant 0$? Use the following graphic interpretation of solutions, which is a modification of the interpretation given in Fig. 4. Let AB be a segment of length $n + r$. Correspond to a solution $(x_1, \dots, x_r)$ the partition of this segment consisting of the segment of length x_1 starting at A, the segment of length x_2, starting at the first integral point after the end of the first segment, etc; see the figure in which we have $x_3 = 0$.

7. Find the number of different partitions $M = M_1 + M_2$ of type (m, m) if $n(M) = 2m$ and the partitions $M = M_1 + M_2$ and $M = M_2 + M_1$ are not taken to be different. The same question for the partitions $M = M_1 + M_2 + M_3$ of type (m, m, m) if $n(M) = 3m$ and if partitions with different order of M_1, M_2, M_3 are not taken to be different. Finally, the same question for the partitions of type (k, k, l, l, l), $n(M) = 2k + 3l$ and partitions in which equivalent subsets have different order are not taken to be different.

8. What is the form of the term of the polynomial $(x_1 + \cdots + x_n)^2$? The same question for the polynomial $(x_1 + \cdots + x_n)^3$.

9. How many terms are there in the polynomial $(x_1 + \cdots + x_r)^n$, supposing that similar terms are grouped together?

10. Express the polynomial $a_1^2 + a_2^2 + \cdots + a_n^2$ in terms of polynomials σ_1 and σ_2. Suppose that the polynomial $x^n + ax^{n-1} + bx^{n-2} + \cdots$ has n real roots. Prove that $a^2 \geqslant 2b$. When does the equality $a^2 = 2b$ take place? Hint: use Bézout's theorem from Section 1 of Chapter II and the fact that a sum of squares of real numbers cannot be negative.

11. Give a combinatorial proof of the relation $C_n^k = C_{n-1}^k + C_{n-1}^{k-1}$ for binomial coefficients (formula (26) of Chapter II), interpreting C_n^k as $v(M, k)$ where $n(M) = n$. Generalize this relation to the numbers $C(n_1, \ldots, n_r)$.

12. Give a combinatorial proof of the relation $C_n^m = C_n^{n-m}$ for binomial coefficients.

(to be continued in the next issue)

I. R. Shafarevich,
Russian Academy of Sciences,
Moscow, Russia

THE TEACHING OF MATHEMATICS
2000, Vol. III, 1, pp. 15–40

SELECTED CHAPTERS FROM ALGEBRA

I. R. Shafarevich

Abstract. This paper is the continuation of the third part of the publication "Selected chapters of algebra", the first two being published in previous issues of the Teaching of Mathematics, Vol. I (1998), 1–22, and Vol. II, 1 (1999), 1–30, and the beginning of this part in Vol. II, 2 (1999), 65–80.

AMS Subject Classification: 00 A 35

Key words and phrases: Algebra of sets, probability theory, Bernoulli's scheme, Chebyshev inequalities.

CHAPTER III. SET (continued)

3. Algebra of sets

If the intersection of two subsets $M_1 \subset M$ and $M_2 \subset M$ is empty (i.e. $M_1 \cap M_2 = \varnothing$), then the union $M_1 \cup M_2$ consists of elements which belong either to M_1 or to M_2, and any element of $M_1 \cup M_2$ can belong only to one of the sets M_1 or M_2. Hence, $M_1 \cup M_2 = M_1 + M_2$, and so $n(M_1 \cup M_2) = n(M_1) + n(M_2)$.

The case when $M_1 \cap M_2$ is not empty can be reduced to the previous one. Denote by M_1' the complement of $M_1 \cap M_2$ with respect to M_1, that is to say the set of those elements of M_1 which do not belong to $M_1 \cap M_2$. Then $M_1 = (M_1 \cap M_2) + M_1'$ and

$$n(M_1) = n(M_1 \cap M_2) + n(M_1'). \tag{17}$$

Analogously,

$$n(M_2) = n(M_1 \cap M_2) + n(M_2'), \tag{18}$$

where M_2' is the complement of $M_1 \cap M_2$ with respect to M_2. Adding up (17) and (18) we obtain

$$n(M_1) + n(M_2) = 2n(M_1 \cap M_2) + n(M_1') + n(M_2'). \tag{19}$$

But the sets $M_1 \cap M_2$, M_1' and M_2' do not have common elements and their union is $M_1 \cup M_2$. Therefore $M_1 \cup M_2 = M_1 \cap M_2 + M_1' + M_2'$ and so $n(M_1 \cup M_2) =$

This paper is an English translation of the second part of: И. Р. Шафаревич, *Избраные главы алгебры*, Математическое образование, 3, окт.–дек. 1997, Москва, стр. 2–45. In the opinion of the editors, the paper merits wider circulation and we are thankful to the author for his kind permission to let us make this version.

$n(M_1 \cap M_2) + n(M_1') + n(M_2')$. Using this, we can rewrite the equality (19) as: $n(M_1) + n(M_2) = n(M_1 \cap M_2) + n(M_1 \cup M_2)$, that is to say

$$n(M_1 \cup M_2) = n(M_1) + n(M_2) - n(M_1 \cap M_2). \tag{20}$$

This is the relation we wanted. Our further aim is to generalize it and to obtain the expression for the number of elements of the union of an arbitrary number of sets $n(M_1 \cup \dots \cup M_r)$, and not only the union of two sets. We shall have to establish some more or less evident properties of intersections and unions of several sets.

First of all notice that the union $M_1 \cup M_2 \cup \dots \cup M_r$ of several subsets M_1, M_2, ... , M_r can be defined by means of unions of only two subsets. For instance,

$$M_1 \cup M_2 \cup M_3 = (M_1 \cup M_2) \cup M_3,$$

and also for arbitrary k

$$M_1 \cup M_2 \cup \dots \cup M_k = (M_1 \cup M_2 \cup \dots \cup M_{k-1}) \cup M_k.$$

The second formula we need has the form

$$(M_1 \cup M_2 \cup \dots \cup M_k) \cap N = (M_1 \cap N) \cup (M_2 \cap N) \cup \dots \cup (M_k \cap N).$$

Both formulas are obvious; it is enough to ask oneself: what does it mean that an element $a \in M$ belongs to the left or to the right-hand side? For example, in the last formula $a \in (M_1 \cup M_2 \cup \dots \cup M_k) \cap N$ means that $a \in M_1 \cup M_2 \cup \dots \cup M_k$ and $a \in N$. The second statement is merely that $a \in N$ and the first that $a \in M_i$ for some $i = 1, \dots, k$. But then $a \in M_i \cap N$ for the same i, and this means that $a \in (M_1 \cap N) \cup (M_2 \cap N) \cup \dots \cup (M_k \cap N)$. Notice that this property resembles the distributivity of numbers. Indeed, if we replace the sets M_1, M_2, ... , M_k and N by the numbers a_1, a_2, ... , a_k and b, if we replace the sign $\cup$ by $+$ and $\cap$ by $\cdot$, we obtain the equality $(a_1 + \dots + a_k)b = a_1 b + \dots + a_k b$, i.e. the distributivity law for numbers. There are other properties which show an analogy between the operations union and intersection of subsets on one side, and addition and multiplication of numbers on the other (see Problem 1). Investigation of system of subsets of a given set M with respect to the operations $\cup$ and $\cap$ is called the *algebra of sets.*

We now derive the formula for $n(M_1 \cup M_2 \cup M_3)$. Since $M_1 \cup M_2 \cup M_3 = (M_1 \cup M_2) \cup M_3$, we can apply formula (20) to obtain

$$n(M_1 \cup M_2 \cup M_3) = n((M_1 \cup M_2) \cup M_3) = n(M_1 \cup M_2) + n(M_3) - n((M_1 \cup M_2) \cap M_3).$$

We can apply formula (20) to the term $n(M_1 \cup M_2)$ and since $(M_1 \cup M_2) \cap M_3 = (M_1 \cap M_3) \cup (M_2 \cap M_3)$, we can also apply formula (20) to the last term above. We get

$$\begin{aligned} n(M_1 \cup M_2 \cup M_3) = n(M_1) + n(M_2) + n(M_3) & \\ - n(M_1 \cap M_2) - n(M_1 \cap M_3) - n(M_2 \cap M_3) & \\ + n((M_1 \cap M_3) \cap (M_2 \cap M_3)). & \end{aligned}$$

Clearly, $(M_1 \cap M_3) \cap (M_2 \cap M_3) = M_1 \cap M_2 \cap M_3$ and so the last term can be written as $n(M_1 \cap M_2 \cap M_3)$. We obtain the formula

$$\begin{aligned} n(M_1 \cup M_2 \cup M_3) = n(M_1) + n(M_2) + n(M_3) \\ - n(M_1 \cap M_2) - n(M_1 \cap M_3) - n(M_2 \cap M_3) \\ + n(M_1 \cap M_2 \cap M_3). \end{aligned}$$

Now we can guess what should be the form of the formula for $n(M_1 \cup \cdots \cup M_r)$. It must contain the terms $n(M_{i_1} \cap \cdots \cap M_{i_k})$ where $M_{i_1}, \ldots, M_{i_k}$ are any k sets taken among the sets $M_1, \ldots, M_r$ for all $k = 1, 2, \ldots, r$ and if k is even we take the sign $-$, while if k is odd we take $+$. In other words, the sign of the term $n(M_{i_1} \cap \cdots \cap M_{i_k})$ is $(-1)^{k-1}$.

We shall now prove this formula by induction on r in the same way as we proved it for $r = 3$. The induction basis will be formula (20). Write $M_1 \cup M_2 \cup \cdots \cup M_r$ in the form $(M_1 \cup M_2 \cup \cdots \cup M_{r-1}) \cup M_r$, and use formula (20):

$$\begin{aligned} n(M_1 \cup M_2 \cup \cdots \cup M_r) = n(M_1 \cup M_2 \cup \cdots \cup M_{r-1}) + n(M_r) \\ - n((M_1 \cup M_2 \cup \cdots \cup M_{r-1}) \cap M_r). \end{aligned}$$

By the induction hypothesis, the formula is true for $n(M_1 \cup \cdots \cup M_{r-1})$ and gives those terms of $n(M_1 \cup \cdots \cup M_r)$ which do not contain M_r. Now we have

$$(M_1 \cup M_2 \cup \cdots \cup M_{r-1}) \cap M_r = (M_1 \cap M_r) \cup (M_2 \cap M_r) \cup \cdots \cup (M_{r-1} \cap M_r)$$

and by the induction hypothesis we can also apply the formula to the expression $n((M_1 \cap M_r) \cup \cdots \cup (M_{r-1} \cap M_r))$. The intersection

$$(M_{i_1} \cap M_r) \cap \cdots \cap (M_{i_k} \cap M_r)$$

is obviously $M_{i_1} \cap \cdots \cap M_{i_k} \cap M_r$ and so we obtain all the terms of the formula which contain M_r. Moreover, if the term of the formula for $n((M_1 \cap M_r) \cup \cdots \cup (M_{r-1} \cap M_r))$ had the sign $(-1)^{r-1}$, it has the sign $(-1)^r$ in the formula for $n(M_1 \cup \cdots \cup M_r)$ and it will depend on $k+1$ sets $M_{i_1} \cap \cdots \cap M_{i_k} \cap M_r$.

The formula for $n(M_1 \cup \cdots \cup M_r)$ can be written down more conveniently if we consider the number of elements of the complement $\overline{M_1 \cup \cdots \cup M_r}$ of the set $M_1 \cup \cdots \cup M_r$, i.e. the number of elements of the set M which do not belong to any of the subsets M_i. Since for any subset $N \subset M$ we always have $M = N + \overline{N}$, then $n(\overline{N}) = n(M) - n(N)$. In our case $n(\overline{M_1 \cup \cdots \cup M_r})$ will be the sum of the terms $(-1)^k n(M_{i_1} \cap \cdots \cap M_{i_k})$, where $M_{i_1}, \ldots, M_{i_k}$ are any k subsets taken from $M_1, \ldots, M_r$. For $k = 0$ we take the term $n(M)$. In other words,

$$\begin{aligned} n(\overline{M_1 \cup \cdots \cup M_r}) = n - n(M_1) - \cdots - n(M_r) \\ + n(M_1 \cap M_2) + \cdots \\ + (-1)^r n(M_1 \cap \cdots \cap M_r), \end{aligned} \tag{21}$$

where $n = n(M)$.

This formula consists of expressions $n(M_{i_1} \cap \dots \cap M_{i_k})$ where $i_1, \dots, i_k$ are any k elements of the set $\{1, 2, \dots, n\}$. We have met such expressions in connection with Viète's formula (formula (12)). It is worthwhile to compare these two formulas. Formula (21) follows from (12) if we put $x = 1$, $a_i = -x_i$ in (12) and then replace everywhere $x_{i_1}, \dots, x_{i_k}$ by $n(M_{i_1} \cap \dots \cap M_{i_k})$. Indeed, it is sometimes written in this "symbolic" way

$$n(\overline{M_1 \cup \dots \cup M_r}) = n(1 - M_1) \cdots (1 - M_r), \tag{22}$$

where we suppose that the product on the right-hand side is expanded by Viète's formula as if M_i were variables, and then the expression $n \cdot M_{i_1} \cdots M_{i_k}$ (which is meaningless) is replaced by the expression $n(M_{i_1} \cap \dots \cap M_{i_k})$, and $n \cdot 1$ is replaced by $n = n(M)$.

Formula (22) can serve as a means for remembering formula (21), but in algebra, whenever two relations, concerned with different questions, have the same form, it is *always* possible to devise such a definition so that one formula coincides with the other. We shall show this on the example of formulas (21), (22) and Viète's formula (12).

In order to do this we shall have to consider functions on a set M. Undoubtedly, you must have already met with the concept of a function—in one way or another. By a function we shall mean any way of corresponding to each element $a \in M$ a certain number. The actual process of corresponding will be denoted by f, and the number corresponded to the element a by this process will be denoted by $f(a)$. It is also called the value of the function f at the element a. Although the concept of a function is defined for arbitrary sets, we shall at the moment be interested only in the case when the set M is finite. Then a function can be represented by writing with each element a its corresponding number $f(a)$. For example, here are two functions f and g, defined on the set of three elements $M = \{a, b, c\}$ (Fig. 5).

Fig. 5

Therefore, if $M = \{a_1, \dots, a_n\}$, then a function on M is the sequence $(f(a_1), \dots, f(a_n))$. Functions can be added up and multiplied, these operations being defined by the values of the functions. In other words, the functions $f + g$ and fg are defined by $(f + g)(a) = f(a) + g(a)$ and $(fg)(a) = f(a)g(a)$ for arbitrary a. For example, if f and g are represented by Fig. 5, then $f + g$ and fg are represented by Fig. 6.

Fig. 6

Since the operations with functions are defined by their values, they have the same properties as the operations with numbers: commutativity, associativity, distributivity, etc. We can apply any identity, proved for numbers, if we replace numbers by functions on a given set M. The function $f_M(a)$ which to any element $a \in M$ corresponds the number 1 is denoted by $\mathbf{1}$. Clearly, $\mathbf{1} \cdot f = f$ for any function f.

We now connect the notion of a function with the notion of a subset. For any subset $N \subset M$ there exists the function defined as follows: the values of elements which belong to N are 1, and the values of those which do not belong to N are 0. This function is called the *characteristic function* of the subset N and is denoted by f_N. Thus, $f_N(a) = 1$ if $a \in N$ and $f_N(a) = 0$ if $a \in \overline{N}$. Conversely, it is clear that the function $f_N(a)$ defines the set N—it consists of all elements $a \in M$ such that $f_N(a) = 1$. (In this way we obtain a one-to-one correspondence between the subsets $N \subset M$ and those functions which take only two values 0, 1. This is the same relation which enables us to deduce Theorem 1 from Theorem 2. See Problem 1 from Section 2, where p and q should be replaced by 0 and 1.)

Some properties of subsets are simply expressed in terms of their characteristic functions. For example, the characteristic function of the whole set M has all values equal to 1, and hence $f_M = \mathbf{1}$. If $\overline{N}$ is the complement of N, then $f_{\overline{N}} = \mathbf{1} - f_N$: indeed, if $a \in N$, i.e. $f_N(a) = 1$, then $(\mathbf{1} - f_N)(a) = 0$, as it should be. Analogously for $a \in \overline{N}$. If N_1 and N_2 are arbitrary subsets then $f_{N_1 \cap N_2} = f_{N_1} \cdot f_{N_2}$, since if $a \in N_1$ and $a \in N_2$ then $f_{N_1} f_{N_2}(a) = 1 \cdot 1 = 1$. If a does not belong to one of the sets N_1 or N_2, then one of the factors f_{N_1} or f_{N_2} is 0 and so $f_{N_1} f_{N_2}(a) = 0$, and also $f_{N_1 \cap N_2}(a) = 0$. Clearly, this is also true for several subsets:

$$\text{(23)} \qquad \text{if} \quad N' = N_1 \cap \cdots \cap N_r, \quad \text{then} \quad f_{N'} = f_{N_1} \cdots f_{N_r}.$$

We can now rewrite formula (21) in the language of characteristic functions. First of all, notice that the considered set $\overline{M_1 \cup \cdots \cup M_r}$ is equal to $\overline{M_1} \cap \cdots \cap \overline{M_r}$. This is evident: an element a does not belong to the set $M_1 \cup \cdots \cup M_r$ if it does not belong to any of M_i, i.e. if it belongs to all $\overline{M_i}$. Now using formula (23) we can write the characteristic function of the set $\overline{M_1 \cup \cdots \cup M_r}$ in the form

$$f_{\overline{M_1 \cup \cdots \cup M_r}} = f_{\overline{M_1} \cap \cdots \cap \overline{M_r}} = f_{\overline{M_1}} \cdots f_{\overline{M_r}}.$$

Besides, we know that $f_{\overline{M_i}} = \mathbf{1} - f_{M_i}$ and we obtain

$$f_{\overline{M_1 \cup \cdots \cup M_r}} = (\mathbf{1} - f_{M_1})(\mathbf{1} - f_{M_2}) \cdots (\mathbf{1} - f_{M_r}).$$

We can now apply Viète's formula (13), by setting into it $x = \mathbf{1}$, $a_i = f_{M_i}$. We have already explained why this is possible. We obtain

$$f_{\overline{M_1 \cup \cdots \cup M_r}} = \mathbf{1} - \sigma_1(f_{M_1}, \dots, f_{M_r}) + \sigma_2(f_{M_1}, \dots, f_{M_r}) \\ - \cdots + (-1)^r \sigma_r(f_{M_1}, \dots, f_{M_r}).$$

Moreover, $\sigma_k(f_{M_1}, \dots, f_{M_r})$ is the sum of all products $f_{M_{i_1}} \cdots f_{M_{i_k}}$ for all different indices $(i_1, \dots, i_k)$ taken from $(1, \dots, n)$. We know that $f_{M_{i_1}} \cdots f_{M_{i_k}} = f_{M_{i_1} \cap \cdots \cap M_{i_k}}$ and we obtain that

$$(24) \qquad f_{\overline{M_1 \cup \cdots \cup M_r}} = \mathbf{1} - f_{M_1} - \cdots - f_{M_r} + f_{M_1 \cap M_2} + \cdots + (-1)^r f_{M_1 \cap \cdots \cap M_r},$$

i.e. the sum of all functions $f_{M_{i_1} \cap \cdots \cap M_{i_k}}$ which are taken with the $+$ sign if k is even and with the $-$ sign if k is odd.

Notice that we have obtained something essentially more than the formula (21): we found the expression not for the number of elements $n(\overline{M_1 \cup \cdots \cup M_r})$ of the subset $\overline{M_1 \cup \cdots \cup M_r}$, but for its characteristic function which does not determine only the number of elements of the subset, but the subset itself. In particular, formula (21) has sense only when the set M is finite, while the relation (24) is true for a finite number of subsets of an arbitrary set M.

In order to deduce the relation (21) from it, we have to return from functions back to numbers. It is essential here that M be finite. For any function we define the number Sf as the sum of all values $f(a)$ of the function f at all elements $a \in M$: if $M = \{a_1, \dots, a_n\}$, then $Sf = f(a_1) + \cdots + f(a_n)$. For example, for the functions f and g from Fig. 5 we have $Sf = 2$, $Sg = 1$, Clearly, for any two functions f, g we have $S(f + g) = Sf + Sg$. Indeed, the value of $f + g$ at a_i is $f(a_i) + g(a_i)$. Therefore, $S(f + g) = (f(a_1) + g(a_1)) + \cdots + (f(a_n) + g(a_n)) = (f(a_1) + \cdots + f(a_n)) + (g(a_1) + \cdots + g(a_n)) = Sf + Sg$. If f_N is the characteristic function of the subset N, then $f_N(a) = 1$ is true for the elements $a \in N$, and for the other elements a it is 0. Hence, $Sf_N = n(N)$.

If we now find the number Sf for the functions on the left and right-hand side of (24), using the established properties, we obtain exactly the relation (21).

Consider now two applications of formula (21). The first is a question studied long time ago by Euler, and it concerns the permutations of a set. We said at the end of the last Section (Remark 1) that this is the name for the arrangements of elements of a set M in a given order. The number of permutations is $n!$ if $n(M) = n$. At the end of Section 2 we wrote down, as an example, all the six permutations of the three element set $M = \{a, b, c\}$. In the general case we also write down all the $n!$ permutations of the set M and denote by $(a_1, \dots, a_n)$ the first one. The question is: how many permutations do we have in which no element takes the same place as in the first one? This is precisely Euler's question. Solve it for the case $n = 3$ and the six permutations written at the end of Section 2. Verify that only two permutations satisfy the given condition, namely: (c, a, b) and (b, c, a).

In the general case we shall apply formula (21). Denote by $\mathcal{P}$ the set of all permutations of the elements of the set $M = \{a_1, \dots, a_n\}$. We have

$n(\mathcal{P}) = n!$. Consider those permutations in which a_i stands at the same place as in the first permutation, i.e. at the i-th place. Denote the set of all such permutations by $\mathcal{P}_i$. Then our question becomes: find $n(\overline{\mathcal{P}_1 \cup \cdots \cup \mathcal{P}_n})$. Hence we have the same situation as we had before for the set $\mathcal{P}$ and its subsets $\mathcal{P}_1, \dots, \mathcal{P}_n$ (in formula (21) the set was denoted by M and its subsets by M_i). In order to apply the formula, we have to find the numbers $n(\mathcal{P}_{i_1} \cap \cdots \cap \mathcal{P}_{i_k})$. But the set $\mathcal{P}_{i_1} \cap \cdots \cap \mathcal{P}_{i_k}$ contains exactly those permutations in which $a_{i_1}, \dots, a_{i_k}$ take the same place as in the first permutation, namely they are the places $i_1, \dots, i_k$, respectively. Such a permutation differs from the first permutation only in the arrangement of elements in other places. In other words, the number of such permutations is equal to the total number of permutations of the set $\overline{\{a_{i_1}, \dots, a_{i_k}\}}$. Since $n(\overline{\{a_{i_1}, \dots, a_{i_k}\}}) = n - k$, applying the general formula we get $n(\mathcal{P}_{i_1} \cap \cdots \cap \mathcal{P}_{i_k}) = (n-k)!$. All the sets $\mathcal{P}_{i_1} \cap \cdots \cap \mathcal{P}_{i_k}$ for a fixed k give one term in the formula (21), and the number of such terms is equal to the number of subsets $\{a_{i_1}, \dots, a_{i_k}\} \subset \{a_1, \dots, a_n\}$ for a given k, that is to say, according to Theorem 3 it is C_n^k. Hence, the contribution of the terms which correspond to a given value of k is $C_n^k (n-k)!$ and substituting the value of the binomial coefficient we get $\dfrac{n!}{k!\,(n-k)!}(n-k)! = \dfrac{n!}{k!}$ and formula (21) in our case becomes

$$\begin{aligned} n(\overline{\mathcal{P}_1 \cup \cdots \cup \mathcal{P}_n}) &= n! - \frac{n!}{1!} + \frac{n!}{2!} - \cdots + (-1)^n \frac{n!}{n!} \\ &= n! \left(1 - \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{(-1)^n}{n!} \right). \end{aligned}$$

This is the formula founded by Euler. He was actually interested in the ratio of the founded number with the number of all permutations $n!$. This ratio is $1 - \dfrac{1}{1!} + \dfrac{1}{2!} + \cdots + \dfrac{(-1)^n}{n!}$, which, as n increases, can be shown to approach a fixed number, namely $1/e$, where e is the basis of natural logarithms (for those who already know what that is). The number $1/e$ is irrational and approximately equal to 0,36787 … .

The second application of formula (21) is related to the properties of positive integers. Let n be a positive integer, and let $p_1, \dots, p_r$ be its prime divisors, different from each other. How many positive integers exist which are not greater than n and which are not divisible by any of the numbers p_i? This is again an application of formula (21). Denote by M the set of positive integers 1, 2, … , n and by M_i its subset whose elements are divisible by p_i. Clearly, our problem is equivalent to the evaluation of $n(\overline{M_1 \cup \cdots \cup M_r})$. Let us find the values of the terms $n(M_{i_1} \cap \cdots \cap M_{i_k})$ in formula (21). The set $M_{i_1} \cap \cdots \cap M_{i_k}$ consists of all positive integers $t \leqslant n$ which are divisible by prime numbers $p_{i_1}, p_{i_2}, \dots, p_{i_k}$. This is equivalent to the fact that t is divisible by their product $p_{i_1} p_{i_2} \dots p_{i_k}$. Let m be a divisor of n. How many are there positive integers $t \leqslant n$ which are divisible by m? Such numbers have the form $t = mu$, where u is a positive integer and the condition $t \leqslant n$ is equivalent to $u \leqslant n/m$. Hence, u may take the values 1, 2, … , n/m and the number of such numbers is n/m. If $m = p_{i_1} \cdots p_{i_k}$ this gives that $n(M_{i_1} \cap \cdots \cap Mi_k) = \dfrac{n}{p_{i_1} \cdots p_{i_k}}$ and formula (21) becomes

$$n(\overline{M_1 \cup \cdots \cup M_r}) = n - \frac{n}{p_1} - \cdots - \frac{n}{p_r} + \frac{n}{p_1 p_2} + \cdots + (-1)^r \frac{n}{p_1 \cdots p_r}.$$

The right-hand side can be written in the form

$$n\left(1 - \frac{1}{p_1} - \frac{1}{p_2} - \cdots + \frac{1}{p_1 p_2} + \cdots + (-1)^r \frac{1}{p_1 \cdots p_r}\right).$$

The expression in brackets can be transformed by Viète's formula (applied simply to numbers), if we set $x = 1$, $\alpha_i = -1/p_i$. By (13) this expression will be

$$\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right).$$

Therefore, for the number of positive integers not greater than n and not divisible by $p_1, p_2, \ldots, p_r$ we obtain

$$n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right). \tag{25}$$

We often meet the case when $p_1, \ldots, p_r$ are all the prime divisors of n. In this case t is not divisible by any of p_i's if and only if it is relatively prime to n: if it had a common factor d with n, then this factor would have a prime divisor p_i which would divide both t and n. Therefore, formula (25) gives the number of all positive integers not greater than n and relatively prime to n, if we take $p_1, \ldots, p_r$ to be all prime divisors of n. The expression (25) was found in this form by Euler, it is denoted by $\varphi(n)$, and is called Euler's function. For example, for $n = 675 = 3^3 \cdot 5^2$ we have $n(1 - \frac{1}{3})(1 - \frac{1}{5}) = 3^2 \cdot 5(3 - 1)(5 - 1) = 360$ numbers which are not greater than 675 and which are relatively prime to 675.

Suppose now that $p_1, \ldots, p_r$ need not necessarily divide n. What is the number of positive integers $t \leqslant n$ which are not divisible by $p_1, \ldots, p_r$? We can repeat the previous reasoning, but with one alternation. We have to find the number of positive integers $t \leqslant n$ divisible by $p_{i_1} \cdots p_{i_r}$. Let m be an arbitrary positive integer. How many are there positive integers $t \leqslant n$ which are divisible by m? Again put $t = mu$ with the condition $mu \leqslant n$. Hence, we have to take all numbers $u = 1, 2, \ldots$, such that $mu \leqslant n$. Let u be the last of them. Then $r = n - mu < m$, for in the opposite case such a number would also be $mu + m = n(u + 1)$. But then $n = mu + r$ where $0 \leqslant r < m$—which is the formula for the division with remainder of n by m (see Theorem 4 of Chapter I). Hence the number u is equal to the quotient in the above division and it shall be denoted by $[n/m]$. Therefore, the number of positive integers not greater than n and divisible by m is $[n/m]$. We can now literally repeat the preceding argument and apply the formula (21). For the number of positive integers not greater than n and not divisible by $p_1, \ldots, p_r$ we obtain the expression

$$n - \left[\frac{n}{p_1}\right] - \left[\frac{n}{p_2}\right] - \cdots + \left[\frac{n}{p_1 p_2}\right] + \cdots + (-1)^r \left[\frac{n}{p_1 \cdots p_r}\right]. \tag{26}$$

It is not as explicit as the expression (25) but we can write it in the form of (25), as an approximation. Recall the formula for the division with a remainder: $n = mu+r$, where $0 \leqslant r < m$ and $u = [n/m]$. Dividing this by m we obtain $\dfrac{n}{m} = u + \dfrac{r}{m}$ and since $0 \leqslant r < m$, we get $\dfrac{n}{m} - 1 < \left[\dfrac{n}{m}\right] \leqslant \dfrac{n}{m}$. In other words, we can replace $[n/m]$ by n/m with an error less than 1. Make this replacement in all the terms of (26). What is the total error? Each term of (26) corresponds to a subset $\{i_1, \ldots, i_k\}$ of the set $\{1, \ldots, r\}$. According to Theorem 2, the number of such subsets is 2^r. Hence this is the number of terms in (26). Since each replacement produces an error less than 1, the total error will be less than 2^r. That is to say that the expression (26) differs from

$$n - \frac{n}{p_1} - \frac{n}{p_2} - \cdots + \frac{n}{p_1 p_2} + \cdots + (-1)^r \frac{n}{p_1 \cdots p_r} \tag{27}$$

by less than 2^r. We have met with the last expression before, and we know that it is equal to

$$n\left(1 - \frac{1}{p_1}\right) \cdots \left(1 - \frac{1}{p_r}\right).$$

In this way we obtain that for the number N of positive integers not greater than n and not divisible by given prime numbers $p_1, \ldots, p_r$ the following inequality holds

$$\left| N - n\left(1 - \frac{1}{p_1}\right) \cdots \left(1 - \frac{1}{p_r}\right) \right| < 2^r. \tag{28}$$

For example, if we have three prime numbers p, q, r then N is equal to $n\left(1 - \dfrac{1}{p}\right)\left(1 - \dfrac{1}{q}\right)\left(1 - \dfrac{1}{r}\right)$ with an error less than 8.

Problems

1. Verify the relations $M_1 \cap \cdots \cap M_k = (M_1 \cap \cdots \cap M_{k-1}) \cap M_k$ and $(M_1 \cap \cdots \cap M_k) \cup N = (M_1 \cup N) \cap \cdots \cap (M_k \cup N)$. The second of these is again analogous to the distribution law for numbers $(a_1 + \cdots + a_k)b = a_1 b + \cdots + a_k b$, but now the role of multiplication is taken by $\cup$ and the role of addition by $\cap$.

2. Verify that for each relation between subsets involving the operations $\cup$ and $\cap$, there exists another relation in which these two operations change places. In order to do this, prove that $\overline{M_1 \cup M_2} = \overline{M}_1 \cap \overline{M}_2$ and $\overline{M_1 \cap M_2} = \overline{M}_1 \cup \overline{M}_2$.

3. How many times does the function $\sin ax$ take the value 0 on the segment from 0 to $2\pi b$, where $0 < a < b$ and a, b are positive integers?

4. For positive integers $a_1, \ldots, a_m$ the expression $\max(a_1, \ldots, a_m)$ denotes the greatest and $\min(a_1, \ldots, a_m)$ the smallest one of them. Let $N = \max(a_1, \ldots, a_n)$. For the set $M = \{1, \ldots, N\}$ define M_i as the subset consisting of those j's for which $a_j < a_i$. Applying formula (21), find the relation between $\max(a_1, \ldots, a_n)$ and $\min(a_{i_1}, \ldots, a_{i_m})$ where $\{a_{i_1}, \ldots, a_{i_m}\}$ is a subset of $\{a_1, \ldots, a_n\}$.

5. Apply formula (21) for the case when $M_i = \overline{\{\alpha_i\}}$. By evaluating directly all the terms which appear in it, obtain the relation

$$n - C_n^1(n-1) + C_n^2(n-2) + \cdots + (-1)^{n-1} C_n^{n-1} \cdot 1 = 0.$$

6. Let M be a finite set and let h be an arbitrary function on M. For a subset $N \subset M$ define the number $S_h(N)$ as the sum of all values $h(a)$, for all $a \in N$. Prove the formula analogous to (21) where $n(M)$ is replaced everywhere by $h(N)$. Hint: multiply the relation (24) by the function h.

7. Find the sum of all positive integers not exceeding n and relatively prime to n. Hint: apply the result of Problem 6 with $h(k) = k$.

8. The same question for the sum of squares of these numbers.

9. Prove that in the right-hand side of inequality (28) we may replace 2^r by 2^{r-1}.

4. The language of probability

The theory of probability, like any other branch of mathematics, has its basic concepts which are not defined—like points or numbers. The first such a concept is the *event*. In this Section we shall consider the case when the number of events is finite. Usually, an event is the result of the occurrences of some simpler events which are said to be *elementary*. For instance, when we throw dice there are 6 possible elementary events: the appearance of number 1, number 2, number 3, number 4, number 5, number 6 on the top face. The event that we obtain an even number consists of three elementary events: either we obtain 2, or 4, or 6. The set of elementary events is simply a set (in this Section a finite set) whose elements have special names (elementary events). An event is a *subset* of the set of elementary events. The second basic concept is *probability*: it is a real number assigned to each elementary event. Therefore, if $M = \{a_1, \dots, a_n\}$ is the set of elementary events, then to define a probability means assigning to each element $a_i \in M$ a real number p_i, which is called the probability of the event a_i. Probabilities should satisfy two conditions: they should be nonnegative and the sum of probabilities of all elementary events should be equal to 1:

$$(29) \qquad p_i \geqslant 0, \qquad p_1 + \cdots + p_n = 1.$$

In other words, probability is a function $p(a)$ on the set of elementary events M with real values, satisfying the conditions $p(a) \geqslant 0$ for all $a \in M$ and the sum of all the numbers $p(a)$ for $a \in M$ is 1. These conditions play the role of axioms of probability. If N is an arbitrary event (recall that an event is a subset of the set M) then its probability is the sum of the numbers $p(a)$ for all $a \in N$. This probability is denoted by $p(N)$. In the special case when $N = M$, the corresponding event is said to be *certain*. The condition (29) shows that the probability of the certain event is 1. The condition $p(M) = 1$ is not as essential as the condition $p(M) > 0$. The arbitrary case can be reduced to the case $p(M) = 1$, by dividing all the probabilities by $p(M)$. We simply choose the probability of the certain event to be the unit of measuring other probabilities. We emphasize that the object studied by the theory of probability is the set (in our case finite) of elementary events with prescribed probabilities. This set and the probabilities are chosen according to the specific conditions of the considered problem. Afterwards, when they are defined, we can evaluate probabilities of other events. That is why the specialists in the theory of probability say that their task is to find probabilities of certain events using probabilities of other events.

If the two events are given—and we recall that they are simply two subsets N_1 and N_2 of the set M—then their union $N_1 \cup N_2$ and intersection $N_1 \cap N_2$ are also events. From the definition it follows that $p(N_1 \cup N_2) \leqslant p(N_1) + p(N_2)$. The strict inequality may take place since in the sum $p(N_1) + p(N_2)$ the term $p(a)$ will appear twice if $a \in N_1 \cap N_2$. In fact we have

$$p(N_1 \cup N_2) = p(N_1) + p(N_2) - p(N_1 \cap N_2).$$

We came across this relation earlier (see Problem 6 of Section 3). In particular, if $N_1 \cap N_2 = \varnothing$, i.e. if N_1 and N_2 do not intersect, then the events N_1 and N_2 are said to be *mutually exclusive.* In that case $p(N_1 \cup N_2) = p(N_1) + p(N_2)$. A particular case is when $N_1 = N$ is an arbitrary subset and $N_2 = \overline{N}$ is its complement. We obtain that $p(N) + p(\overline{N}) = 1$ or $p(\overline{N}) = 1 - p(N)$. The event $\overline{N}$ is said to be *opposite* to N.

The basic object: the set M and the given function on M satisfying the axioms of probability (29) is called a *probability scheme.* It is denoted by $(M; p)$.

An important case of defining probability schemes is when all the elements of the set M have the same probability, i.e. when all the numbers p_i are equal. From the condition (29) it follows that all p_i's are equal to $1/n$. If $N \subset M$ is an arbitrary event then $p(N) = n(N)/n$. For example, this is the case when we throw dice, if the dice is considered to be homogeneous. In this case, all 6 elementary events which correspond to the possible appearances of the numbers 1, 2, ... , 6 on the top face have the same probability 1/6, and the event that an even number appears on the top face has probability $3 \cdot 1/6 = 1/2$.

If the dice is not homogeneous, we have no reason to give all elementary events equal probabilities. In this case we may define the probabilities experimentally, by throwing dice many times and noting the result. If after a large number n of throwing the number i appears k_i times, then the probability of the elementary event—the appearance of the number i—is taken to be k_i/n. Clearly, the conditions (29) will be satisfied. The number n depends on the accuracy we wish to attain. This gives another probability scheme $(M; p)$.

Analogous to the case of dice throwing is the popular problem of drawing balls from a bag. Suppose that the bag contains n identical balls and that we draw out one of them without looking. The drawing out of a ball is an elementary event. The phrase "identical balls" mathematically means that the probabilities of these events are equal. Hence, they are equal to $1/n$. Suppose now that in the bag we have balls of different colours: a black, and b white balls, where $a + b = n$. Then the event "a white ball is drawn from the bag" is a subset $N \subset M$. Since $n(N) = b$, we have $p(N) = b/n$—this is the probability that a white ball is drawn out.

Somewhat more involved is the dice problem, if dice is thrown twice. In this case an elementary event will be given by two numbers (a, b), where $1 \leqslant a \leqslant 6$, $1 \leqslant b \leqslant 6$ which show that in the first throwing we get a, and in the second b. The number of elementary events is 36. This can be represented by the Table 3, where on the horizontal we write all the possible outcomes of the first throwing, and on the vertical the outcomes of the second. For example, to the elementary event that the first throwing gives 5 and the second 4 corresponds the cell marked with an asterisk. The event that the first throwing gives 5 again has the probability 1/6. But it is no longer elementary: it is comprised of six elementary events which correspond to the cells of the vertical column above the number 5. They correspond to the appearance of any number i, $1 \leqslant i \leqslant 6$ on the top face of the dice in the second throwing, if 5 appeared in the first. Since the first throwing has no effect on the second, and the dice is supposed to be homogeneous, we conclude that all the six elementary events have equal probabilities, and since the probability of the event which they make is 1/6, then the probability of each one of them must be 1/36. Hence, we see that the probability of any elementary event is 1/36.

Table 3

Table 4

Consider the event N_k: "the sum of the numbers obtained at the first and the second throwing is equal to k" ("the score is k"). For each pair (a, b) write in the corresponding cell the sum $a+b$ (Table 4). We see that 12 appears in one cell and so $n(N_{12}) = 1$ and also $n(N_{11}) = 2$, $n(N_{10}) = 3$, $n(N_9) = 4$, $n(N_8) = 5$, $n(N_7) = 6$, $n(N_6) = 5$, $n(N_5) = 4$, $n(N_4) = 3$, $n(N_3) = 2$, $n(N_2) = 1$. The greatest value has $n(N_7)$, and since $p(N_k) = n(N_k)/36$, we see that $p(N_7)$ has the greatest value among all $p(N_k)$'s. In other words, the event that the score 7 will be obtained in two throwing is the most probable.

And what is the answer in the case of n throwing? Here the elementary events are given by sequences of n numbers $(a_1, \dots, a_n)$ where each one can take the values 1, ... , 6. The same reasoning as before shows that their probabilities are $1/6^n$. The event N_k: "the total score after n throwing is k" consists of those sequences which satisfy $a_1 + \cdots + a_n = k$. Hence, we have to find which number k has the greatest number of representations of the form

$$k = a_1 + \cdots + a_n, \qquad 1 \leqslant a_i \leqslant 6. \tag{30}$$

In order to do this, consider the polynomial $F(x) = (x + x^2 + \cdots + x^6)^n$. Expanding it, we take from the i-th bracket the term x^{a_i} and as the result we obtain the term $x^{a_1+\cdots+a_n}$. There are several terms of this form, and we collect them together. Therefore, the number of various representations (30) is equal to the coefficient of x^k in the polynomial $F(x)$, and our problem reduces to finding which term has the greatest coefficient. Since $F(x) = x^n G(x)$, where $G(x) = (1 + x + \dots + x^5)^n$, the coefficient of x^k in $F(x)$ is equal to the coefficient of x^{k-n} in $G(x)$, and it is enough to find the term with the greatest coefficient in $G(x)$.

Polynomial $G(x)$ has two properties from which the answer to the above question follows.

An arbitrary polynomial $f(x) = c_0 + c_1 x + \dots + c_n x^n$ is said to be *reciprocal* if its terms, equidistant from its ends, have equal coefficients, i.e. if $c_k = c_{n-k}$. If the coefficients c_i are represented by points with coordinates (i, c_i) in the plane, this property means that these points will be arranged symmetrically with respect to the middle: the line $x = n/2$. On Fig. 7a) we represent the case when n is even, and on Fig. 7b) the case when n is odd.

a) b)

Fig. 7

The polynomial $x^n f\Big(\dfrac{1}{x}\Big)$ has the same coefficients as the polynomial $f(x)$, but in the reversed order. Indeed, if $f(x) = a_0 + a_1 x + \dots + a_n x^n$, then $f\Big(\dfrac{1}{x}\Big) = a_0 + a_1 \dfrac{1}{x} + \dots + a_n \dfrac{1}{x^n}$ and $x^n f\Big(\dfrac{1}{x}\Big) = a_0 x^n + a_1 x^{n-1} + \dots + a_n$. Therefore, the fact that $f(x)$ is a reciprocal polynomial, means that $x^n f\Big(\dfrac{1}{x}\Big) = f(x)$. This implies that the product of two reciprocal polynomials is also reciprocal. Indeed, if $f(x)$ and $g(x)$ are reciprocal polynomials of degree n and m, then $x^n f\Big(\dfrac{1}{x}\Big) = f(x)$, $x^m g\Big(\dfrac{1}{x}\Big) = g(x)$. Multiplying these equalities we get $x^n f\Big(\dfrac{1}{x}\Big) x^m g\Big(\dfrac{1}{x}\Big) = f(x)g(x)$, i.e. $x^{n+m} f\Big(\dfrac{1}{x}\Big) g\Big(\dfrac{1}{x}\Big) = f(x)g(x)$, which means that the polynomial $f(x)g(x)$ is reciprocal. By induction we conclude that the product of any number of reciprocal polynomials is also reciprocal. Finally, since the polynomial $1 + x + \dots + x^5$ is reciprocal, so is the polynomial $G(x) = (1 + x + \dots + x^5)^n$.

The polynomial $f(x) = c_0 + c_1 x + \dots + c_n x^n$ is called *unimodal* if for some $m \leqslant n$ the following inequalities hold: $c_0 \leqslant c_1 \leqslant \ldots \leqslant c_m \geqslant c_{m+1} \geqslant \ldots \geqslant c_n$. That is to say, the coefficients c_i at first do not decrease, and from a certain moment they do not increase. If they are again represented by the points (i, c_i) then they will have “one hump” (Fig. 8).

Fig. 8

For example, polynomial $(1+x)^n$ is reciprocal: this follows from the property $C_n^m = C_n^{n-m}$ of binomial coefficients (see Section 3 of Chapter II). It is also unimodal: this follows from the property of binomial coefficients proved in Section 3 of Chapter II.

It can be proved that if the polynomials $f(x)$ and $g(x)$ have nonnegative coefficients, if they are reciprocal and unimodal, then $f(x)g(x)$ is unimodal. The proof is quite elementary, but a little involved. From this theorem it follows that the polynomial $G(x)$ is unimodal. However, you can easily prove yourself this special case (Problem 3). Now, it is easy to determine the term with the greatest coefficient in a reciprocal unimodal polynomial. Namely, if the term $c_k x^k$ has the greatest coefficient, since the polynomial is reciprocal we have $c_{n-k} = c_k$ and there is the symmetric term $c_k x^{n-k}$. We can take that $k \leqslant n/2$ and $n-k \geqslant n/2$. Since the polynomial is unimodal, none of the terms $c_i x^i$ where $k \leqslant i \leqslant n-k$ can have smaller coefficient, for otherwise there would be two “humps” on the graph. Hence, the greatest coefficient must be the middle coefficient $c_{n/2}$ if n is even or two “equally middle” coefficients $c_{\frac{n-1}{2}} = c_{\frac{n+1}{2}}$ if n is odd (though there may be other coefficients equal to them). In particular, we see that if n is even, then in $G(x)$ the term $x^{\frac{5n}{2}}$ has the greatest coefficient, and if n is odd then there are two terms of $G(x)$, $x^{\frac{5n-1}{2}}$ and $x^{\frac{5n+1}{2}}$ with equal greatest coefficients.

In the polynomial $F(x)$ this term is multiplied by x^n and has degree $\dfrac{5n}{2} + n = \dfrac{7n}{2}$ if n is even. If n is odd, there are two terms with equal coefficients with degree $\dfrac{5n-1}{2} + n = \dfrac{7n-1}{2}$ and $\dfrac{5n+1}{2} + n = \dfrac{7n+1}{2}$. Therefore, if dice is thrown n times the most probable score is $\dfrac{7n}{2}$ if n is even, and if n is odd there are two scores which are both most probable: $\dfrac{7n-1}{2}$ and $\dfrac{7n+1}{2}$.

Consider one more problem of the same type. A certain quantity of m physical particles are registered by n instruments, so that each particle can be registered by any instrument, and the registration of a particle by all instruments are taken to be equally probable. What is the probability that all instruments register at least one particle? An elementary event here is the registration of a particle by an instrument. Let the instruments be denoted by the elements a of the set M. We have $n(M) = n$. Numerate the particles by 1, 2, $\dots$, m. Then an elementary event is the sequence $(a_1, \dots, a_m)$ where $a_i \in M$ and this sequence indicates that the i-th particle is registered by the instrument a_i. In other words, the set of elementary events is M^m in the sense of the definition given in Section 1. The condition of the problem states that all elementary events have equal probabilities. Since by Theorem 1, $n(M^m) = n^m$, the probability of each elementary event is $1/n^m$. We are interested in the subset $N \subset M^m$ which contains the sequences $(a_1, \dots, a_m)$ in which all the elements of M appear. For example, if $M = \{a, b, c\}$, $m = 4$, then $(a, b, c, a) \in N$, but the sequence (a, b, a, b) does not belong to N, since it does not contain c. Our problem is to find $n(N)$.

Denote by M_a the subset of M^m which consists of the sequences $(a_1, \dots, a_m)$ in which none of the a_i's is equal to a. Then clearly $N = \overline{\bigcup M_a}$, i.e. N is the complement of the union of all sets M_a for all $a \in M$. Therefore, $n(N) = n(M^m) - n(\bigcup M_a)$ and the values of the numbers $n(\bigcup M_a)$ are given by formula (21). Let us find the number $n(M_{a_1} \cap M_{a_2} \cap \dots \cap M_{a_r})$, where a_1, $\dots$, a_r are different elements of the set M. Hence, we are dealing with the sequences $(c_1, \dots, c_m)$ in which none of the c_i's equals any of a_1, $\dots$, a_r. In other words, c_i are arbitrary elements of the set $\overline{\{a_1, \dots, a_r\}}$, where $\overline{\{a_1, \dots, a_r\}}$ is the complement of $\{a_1, \dots, a_r\}$ with respect to M. The set of all such sequences is the set $(\overline{\{a_1, \dots, a_r\}})^m$ and the number of elements of this set is, by Theorem 1, $(n(\overline{\{a_1, \dots, a_r\}}))^m$. Since $n(\{a_1, \dots, a_r\}) = r$, $n(M) = n$, we have $n(\overline{\{a_1, \dots, a_r\}}) = n-r$ and $n(M_{a_1} \cap M_{a_2} \cap \dots \cap M_{a_r}) = (n-r)^m$. Hence, each term $n(M_{i_1} \cap M_{i_2} \cap \dots \cap M_{i_r})$ in formula (21) in our case is $(n - r)^m$. The number of terms for a given r is equal to C_n^r, as we know. Therefore, formula (21) gives

$$n(\bigcup M_a) = C_n^1 (n-1)^m - C_n^2 (n-2)^m + \dots + (-1)^n C_n^{n-1} \cdot 1^m.$$

For $N = \overline{\bigcup M_a}$ we obtain

$$n(N) = n(M^m) - n(\bigcup M_a) = n^m - C_n^1 (n-1)^m + \dots + (-1)^{n-1} C_n^{n-1} \cdot 1^m.$$

The requested probability is

$$\frac{n(N)}{n^m} = 1 - C_n^1 \left(\frac{n-1}{n}\right)^m + \dots + (-1)^{n-1} C_n^{n-1} \left(\frac{1}{n}\right)^m . \tag{31}$$

In all the previous examples elementary events had equal probabilities $1/n$, where n is the number of elementary events. As a result, the evaluation of probabilities of other events reduced to the counting of the number of subsets—i.e. to a problem of combinatorics. We shall now consider examples which are more characteristic for the theory of probability.

Let (M, p) and (N, q) be two probability schemes. Suppose that they are defined by many times repeated experiments—different experiment for each scheme.

The experiment used to define the probability scheme (M, p) will be called experiment A, and the one used for the scheme (N, q) will be called experiment B. Consider now the experiment consisting of consecutive experiments A and B, and let us try to use it to define a new probability scheme. A similar situation was encountered in connection with consecutive throwing dice (see Table 3). Let $n(M) = m$, $M = \{a_1, \dots, a_m\}$, $p(a_i) = p_i$, $n(N) = n$, $N = \{b_1, \dots, b_n\}$, $p(b_i) = q_i$. Then the new experiment defines the following elementary events: in the first experiment we have the event $a \in M$ and in the second $b \in N$. Hence, new elementary events correspond to the pairs (a, b), where $a \in M$, $b \in N$, or to elements of the set $X = M \times N$. What probabilities can be assigned to these elements? They can be reasonably defined if we introduce one more supposition. We shall take that the experiments A and B, used to define probability schemes (M, p) and (N, q) are *independent*. This means that the result of the second experiment (i.e. B) does not depend on the outcome of the first experiment (i.e. A). Using this condition it is possible to define the probabilities $p(a, b)$ of the events (a, b). Our reasoning will closely follow the one applied in connection with throwing dice two times (see Table 3).

As in that case (and as we did in Section 1), we represent the elements of the set in the form of the rectangular table

Table. 5

The event that the event a_i takes place in the first experiment has, by condition, probability p_i. It is not an elementary event, since it consists of elementary events (a_i, b_1), (a_i, b_2), … , (a_i, b_n), displayed in the i-th column of Table 5. As we agreed the probabilities of these events should not depend on the experiment A, but should be like the probabilities of b_1, … , b_n in the scheme (N, q). But here we arrive at a contradiction: the sum of probabilities of the events (a_i, b_1), (a_i, b_2), … , (a_i, b_n) is equal to p_i, and the sum of the probabilities of the events b_1, … , b_n is 1. In other words, the i-th column is itself a probability scheme, which must be "the same" as the scheme (N, q). But in this scheme the condition (29) is not fulfilled. We therefore have to make the following "correction": we divide the probabilities of all elementary events by the probability of the events p_i. We therefore obtain the probability scheme with probabilities $\dfrac{p((a_i, b_j))}{p_i}$. Since it should coincide with the probability scheme (N, q), we arrive at the equality $\dfrac{p((a_i, b_j))}{p_i} = q_j$, i.e. $p((a_i, b_j)) = p_i q_j$. Hence, we have, *by definition*

$$p(a_i, b_j) = p_i q_j. \tag{32}$$

In this way we obtain the new probability scheme: the sum of probabilities of elementary events which appear in the i-th column of Table 5 is equal to $p_i q_1 + \cdots + p_i q_n = p_i(q_1 + \cdots + q_n) = p_i$, and the sum of the probabilities of all elementary events is $p_1 + \cdots + p_m = 1$. Hence, the condition (29) is fulfilled.

This new probability scheme (X, r) is called the *product* of probability schemes (M, p) and (N, q). We can write it as follows: if the given schemes are (M, p) and (N, q), then $X = M \times N$ and $p((a, b)) = p(a)q(b)$. The product of probability schemes corresponds to the intuitive idea of the probability scheme defined by two consecutive experiments, *independent* from each other. The above reasoning was necessary to *explain* the motivation for the given definition. Formally, the *definition* is given by the simple equality (32).

Now for several probability schemes $(M_1, p_1), \dots, (M_r, p_r)$ we define the product by induction

$$M_1 \times \cdots \times M_r = (M_1 \times \cdots \times M_{r-1}) \times M_r, \tag{33}$$

where $M_1 \times \cdots \times M_{r-1}$ is taken to be known, and the product of two schemes $M_1 \times \cdots \times M_{r-1}$ and M_r is defined above. Let us decipher this definition. As a set, $M_1 \times \cdots \times M_r$ is the product of sets M_1, M_2, ... , M_r, defined in Section 1. Hence, it consists of arbitrary sequences $(a_1, \dots, a_r)$ where a_i can be any element of M_i. The probability of the elementary event $(a_1, \dots, a_r)$ is

$$p((a_1, \dots, a_r)) = p_1(a_1)p_2(a_2)\cdots p_r(a_r). \tag{34}$$

This can also be verified by induction on r. Indeed, according to definitions (33) and (32), we have $p((a_1, \dots, a_r)) = p(((a_1, \dots, a_{r-1}), a_r)) = p((a_1, \dots, a_{r-1}))p(a_r)$ and by induction hypothesis $p((a_1, \dots, a_{r-1})) = p_1(a_1)p_2(a_2)\cdots p_{r-1}(a_{r-1})$ and this implies (34). This equality can be described as follows: in the sequence $(a_1, \dots, a_r)$ replace each element by its probability and multiply the obtained numbers. This is the probability of the sequence.

We now apply the general construction to the special case of the probability scheme I^n where $i = \{a, b\}$ is the probability scheme consisting of two elementary events with probabilities $p(a) = p$, $p(b) = q$, with necessary conditions $p \geqslant 0$, $q \geqslant 0$, $p + q = 1$. We have already defined I^n as a set in Section 1. It consists of all possible "words" of the type $(a, a, b, b, b, a, b, \dots)$ in the "alphabet" of two letters: a and b. Hence, they will be the elementary events. Their probabilities are defined, according to the above reasoning, as follows: if the letter a appears in the "word" k times and the letter b appears $n - k$ times, then its probability is $p^k q^{n-k}$. Such a probability scheme is called *Bernoulli's scheme.* As we saw, it gives the probabilities of the events a and b in n times repeated experiment, when in each one the event a has probability p and the event b has probability q. Besides, we suppose that the outcome of an experiment does not affect the outcomes of later experiments.

For example, for $n = 3$ we have 8 elementary events (a, a, a), (a, a, b), (a, b, a), (a, b, b), (b, a, a), (b, a, b), (b, b, a), (b, b, b). Their respective probabilities are p^3, p^2q, p^2q, pq^2, p^2q, pq^2, pq^2, q^3. Notice that here the letter p does not denote the probability, but a fixed number, where $0 < p < 1$. The probability of the elementary event which corresponds to the sequence having k letters a and $n - k$ letters b is $p^k q^{n-k}$. These notations are too standard to be changed, but we have to pay attention to what the letter p denotes.

Let us find the probability of the event A_k which consists of a series of n experiments in which the event a occurs k times. These events consist of elementary events given by "words" $(b, a, b, b, b, a, a, \dots)$ in which a appears in exactly k places. The remaining $n - k$ places are occupied by b. By the general formula, such an elementary event has probability $p^k q^{n-k}$. Now how many elementary events make up the event A_k? This is the number of ways in which k elements can be chosen among n indices $1, 2, \dots, n$, i.e. the number of subsets with k elements of a set of n elements. According to Theorem 3, this is the binomial coefficient C_n^k. Therefore for the probability of the event A_k we obtain

$$p(A_k) = C_n^k p^k q^{n-k} = \frac{n!}{k!\,(n-k)!}\, p^k q^{n-k}. \tag{35}$$

Using this we can find the most probable number of occurrences of the event a. It is the value of k for which the expression in (35) has the greatest value. Write down the expressions (35):

$$1 \cdot q^n, \quad npq^{n-1}, \quad \frac{n(n-1)}{2} p^2 q^{n-2}, \quad \dots, \quad 1 \cdot p^n$$

and consider the ratio of two neighbouring terms:

$$\frac{p(A_{k+1})}{p(A_k)} = \frac{n!}{(k+1)!\,(n-k-1)!}\, p^{k+1} q^{n-k-1} \Big/ \frac{n!}{k!\,(n-k)!}\, p^k q^{n-k} = \frac{(n-k)p}{(k+1)q}$$

(after the cancellations, which you can easily check).

If this ratio is greater than 1, then the $(k+1)$-st number is greater than the k-th; if it is 1, then the two numbers are equal, and if it is less than 1, then the $(k+1)$-st number is less than the k-th. The ratio will be greater than 1 if $\dfrac{(n-k)p}{(k+1)q} > 1$, i.e. $(n-k)p > (k+1)q$ or $np > k(p+q)+q$. Having in mind that $p+q = 1$, we can write this inequality in the form $np > k + 1 - p$, i.e. $(n+1)p - 1 > k$. If $k > (n+1)p - 1$, then the ratio $p(A_{k+1})/p(A_k)$ is smaller than 1. Finally, if $k = (n+1)p - 1$, then $p(A_{k+1}) = p(A_k)$. Therefore, as k takes values less than $(n+1)p - 1$, as we move from the k-th number to the $(k+1)$-st, we obtain greater numbers. We distinguish between two cases.

a) The number $(n+1)p-1$ is not an integer. Then the greatest number $p(A_m)$ is obtained for the greatest integer m which does not exceed $(n+1)p$. Moreover, $m \neq (n+1)p - 1$ and for greater values of k such number $p(A_k)$ is smaller than the preceding one. Therefore, there is one most probable number of occurrences of the event a—that is the greatest integer m which does not exceed $(n+1)p - 1$.

b) The number $(n+1)p-1$ is an integer. Then the number $p(A_k)$ increases if $k < m = (n+1)p-1$. Further, $p(A_{m+1}) = p(A_m)$ and for $k > m+1$ the numbers $p(A_k)$ decrease. Hence, the numbers $p(A_k)$ increase until they reach a maximum, then we have one or two numbers equal to this maximum, and then they decrease. In other words, they have "one hump" as in Fig. 8. This means that the polynomial generated by them, namely $q^n+np^{n-1}qt+\frac{n(n-1)}{2}p^{n-2}q^2t^2+\cdots+p^nt^n$ is unimodal. Using the binomial formula, we can write this polynomial in the form $(q+pt)^n$. How can one detect that it is unimodal when it is written in such a simple form? I do not know that such a method exists.

In the simplest case, when $p = q = \frac{1}{2}$, we obtain that if $(n+1)\frac{1}{2}-1$ is not an integer, i.e. if n is even, then $(n+1)\frac{1}{2}-1 = \frac{n}{2}-\frac{1}{2}$ and $m = \frac{n}{2}$. Therefore, there is one most probable number of occurrences of the event a—this is $m = \frac{n}{2}$. This means that it is most likely that both events a and b occur n times. It is not surprising, since such an answer is suggested by symmetry. If n is odd, then $m = (n+1)\frac{1}{2}-1 = \frac{n-1}{2}$ is an integer, and there are two most probable occurrences of the event a: $\frac{n-1}{2}$ (in which case b occurs $\frac{n+1}{2}$ times) and $\frac{n+1}{2}$ (in which case b occurs $\frac{n-1}{2}$ times) and this is also quite natural. But for all other values of p we obtain the answer which would be difficult to predict. Here is a problem from a textbook on probability.

After many years of observations it was concluded that the probability that it will rain on the July 1st is 4/17. Find the most probable number of rainy July 1st's in the next 50 years. We have $n = 50$, $p = 4/17$, $m = (n+1)p-1 = 11$. Hence, the most probable numbers of rainy July 1st's are 11 and 12 (with equal probabilities).

The values of probabilities $C_n^k p^k(1-p)^{n-k}$, $k = 0, 1, \ldots, n$ have many important properties. On Fig. 9, taken from a course of the theory of probability, these values are represented for the cases $p = 1/3$, $n = 4$, 9, 16, 36 and 100.

You see that as n increases, they are not arranged chaotically, but rather they approach a smooth curve. In order to see this better, modify each figure as follows: move the greatest number to the y-axis, decrease the distance between the points on the x-axis (this change of scale was done in Fig. 9) and finally decrease all the numbers proportionally with respect to the greatest number. After this, it turns out that as n increases our points more and more closely approach a certain curve—namely, the graph of the function $y = \frac{1}{\sqrt{2\pi}}c^{x^2}$, where π is the ordinary ratio of the perimeter and the diameter of a circle and c (for those who know that e is the basis of natural logarithms) is equal to $1/\sqrt{e}$.

This assertion, called *Laplace's theorem*, gives, in essence, a more subtle property of binomial coefficients. But in order to prove this theorem we would have to explain the phrase "approaches more and more closely", i.e. we would have to

Fig. 9

introduce the concept of the limit, and we shall not go into this.

Problems

1. In an arbitrary probability scheme (M, p) consider k events: $M_1 \subset M, \dots, M_k \subset M$. Express the probability $p(M_1 \cup M_2 \cup \dots \cup M_k)$ of the event $M_1 \cup M_2 \cup \dots \cup M_k$ in terms of the probabilities $p(M_{i_1} \cap \dots \cap M_{i_r})$ of the events $M_{i_1} \cap \dots \cap M_{i_r}$.

2. Prove that if the polynomial $f(x)$ is reciprocal and unimodal, then the polynomial $f(x)(1 + x)$ also has these two properties.

3. Prove that if the polynomial $f(x)$ is reciprocal and unimodal, then so is the polynomial $f(x)(1 + x + x^2 + x^3 + x^4 + x^5)$. Deduce then that the polynomial $(1 + x + x^2 + x^3 + x^4 + x^5)^n$ is unimodal.

4. Verify that the answer to the problem of m particles and n instruments is $\dfrac{n!}{n^n}$ if $n = m$. What relation between binomial coefficients is obtained if the formula (32) is applied in this case?

5. There are n identical balls in a bag, m white and $n - m$ black balls. We draw out at random r balls. What is the probability that we draw k white and $r - k$ black balls? Hint: "at random" means that the probabilities of any draws of r balls are equal.

6. Prove that if the probability p in Bernoulli's scheme is an irrational number, then there exists exactly one most probable number of occurrences of the event a.

7. The ratio of the most probable number of occurrences of an event a in Bernoulli's scheme and the number n is called the *most probable section.* Prove that, as the number n increases indefinitely, then the most probable section approaches more and more closely the probability p of the event a.

APPENDIX

Inequalities of Chebyshev

We shall again consider a question regarding Bernoulli's scheme which was treated at the end of Section 4. As we said there, Bernoulli's scheme practically arises in the situation when we have several times repeated experiment which can have only two outcomes. For example, suppose that we have an asymmetric (non-homogeneous) coin. The question we pose is: if this coin is spun onto the ground will the top face be "head" or "tail"? In order to arrive at an answer, we make a large series of spins—say, 1000—and if "head" appears k times, we say that the probability of its appearance is $p = k/1000$. After that we can apply our definition of Bernoulli's scheme (I^n, p) and we can find other probabilities within that scheme, e.g. formula (35). But is our abstraction satisfactory? Does it represent sufficiently accurately the reality with which we started: a long series of independent spins? In our abstraction—Bernoulli's scheme—we cannot ask: how many times will the event a occur in the scheme I^n? Since we only operate with the language of probability, we can only pose questions regarding certain probabilities. But the concept of probability is connected with reality by our conviction that an event which has a very small probability practically does not occur. In other words, if the probability of a certain event is sufficiently small, we can in practice proceed as if we knew that it will not occur. Of course, the sense of the words "sufficiently small" has to be made precise in each concrete situation. According to this, we can fix a certain number $\varepsilon > 0$ and consider the following event A_ε: in our Bernoulli's scheme (I^n, p) the event a occurred k times where $\left|\dfrac{k}{n} - p\right| > \varepsilon$. That is to say, the occurrence of the event A_ε means that the "frequency" k/n of occurrences of the event a differs from the supposed probability by more than ε. It is natural to expect that for a fixed ε the probability $p(A_\varepsilon)$ of the event A_ε will become smaller and smaller as n increases indefinitely. It would mean that the difference between the "frequency" k/n and the probability p for large n can be ignored. Jacob Bernoulli considered this problem already at the beginning of the 18th century, and he realized that finding the probability $p(A_\varepsilon)$ is a purely mathematical problem connected with the properties of binomial coefficients. He proved that the probability $p(A_\varepsilon)$ indeed becomes sufficiently small as n increases. In the 19th century Chebyshev proved not only this particular Bernoulli's statement, but he also found a simple explicit inequality for the probability $p(A_\varepsilon)$. We shall expose here his theorem. This is the first time in our text that we come across the work of a Russian mathematician. P. L. Chebyshev lived in the period from 1821 till 1894, and was the founder of the Petersburg mathematical school.

Let us write now in the form of an algebraic formula the expression we are investigating. In Section 4 we considered Bernoulli's scheme (I^n, p) and we found the probability of the event A_k which takes place if in the series of experiments the event a, for which $p(a) = p$ occurs k times (formula (35)):

$$p(A_k) = C_n^k p^k q^{n-k}. \tag{1}$$

We now have a given number ε and we are interested in the event A_ε which takes place if an event A_k with index k occurs where k satisfies the inequality $\left|\frac{k}{n} - p\right| > \varepsilon$. We want to find the probability $p(A_\varepsilon)$ of the event A_ε. Recall that an event (in particular, A_k or A_ε) is a subset of the set I^n. It is clear that the subsets A_k with different indices do not intersect and that A_ε is the union of all subsets A_k for those k's for which $\left|\frac{k}{n} - p\right| > \varepsilon$. Therefore, the probability $p(A_\varepsilon)$ is the sum of probabilities $p(A_k)$ with such indices k. Since $p(A_k)$ is given by (1), this means that we have obtained an explicit, although a bit complicated, expression for the probability $p(A_\varepsilon)$ of the event A_ε. It is more convenient to write the condition $\left|\frac{k}{n} - p\right| > \varepsilon$, which defines our indices k in the equivalent form

$$|k - np| > \varepsilon n. \tag{2}$$

In this way we arrive at the sum

(3)

S_ε – the sum of all expressions $C_n^k p^k q^{n-k}$ for all k, $1 \leqslant k \leqslant n$, satisfying (2).

We see that the probability $p(A_\varepsilon)$ of the event A_ε is equal to S_ε.

Now we can formulate Chebyshev's theorem.

Chebyshev's theorem. *For the probability $p(A_\varepsilon)$ of the event A_ε that the number of occurrences of k events a in Bernoulli's scheme (I^n, p), satisfying the condition $\left|\frac{k}{n} - p\right| > \varepsilon$, the following inequality holds*

$$p(A_\varepsilon) < \frac{pq}{\varepsilon^2 n}. \tag{4}$$

The inequality (4) is sometimes written in the form

$$p\left(\left|\frac{k}{n} - p\right| > \varepsilon\right) < \frac{pq}{\varepsilon^2 n}.$$

It is clear that for given p ($q = 1 - p$) and ε, the right-hand side of the inequality (4) decreases as n increases, which is what we wanted to prove. This particular result is called *Bernoulli's theorem.*

As we saw, the probability $p(A_\varepsilon)$ is equal to the S_ε, defined by (3), and so inequality (4) is equivalent to the inequality

$$S_\varepsilon < \frac{pq}{\varepsilon^2 n}.$$

The proof of Chebyshev's theorem is based upon explicit evaluation of certain sums which we formulate in the form of a lemma.

Lemma *For the probabilities $p(A_k)$, defined by the relation (1), we have*

$$p(A_0)+p(A_1)+p(A_2)+\cdots+p(A_n)=1 \tag{5}$$

$$p(A_1)+2p(A_2)+3p(A_3)+\cdots+np(A_n)=np \tag{6}$$

$$p(A_1)+2^2p(A_2)+3^2p(A_3)+\cdots+n^2p(A_n)=n^2p^2+npq. \tag{7}$$

Proof. Denote the left-hand sides of the equalities (5), (6) and (7) by σ_0, σ_1 and σ_2, respectively. We have already seen in Section 4 that, according to the binomial formula, the probabilities $p(A_k)$ are the coefficients of the polynomial $(pt+q)^n$. That is to say, if we put

$$p(A_0)+p(A_1)t+\cdots+p(A_n)t^n=f(t), \tag{8}$$

then

$$f(t)=(pt+q)^n. \tag{9}$$

Setting $t=1$ into (8) and (9), and using the fact that $p+q=1$, we obtain $\sigma_0=1$, i.e. equality (5).

Consider the derivative $f'(t)$ of the polynomial $f(t)$. From the formula (9), using the rule (19) of Section 2 of Chapter II we obtain that

$$f'(t)=np(pt+q)^{n-1} \tag{10}$$

since $(pt+q)'=p$, by formula (15) of Chapter II. On the other hand, applying formula (15) of Chapter II for $f'(t)$ to the polynomial $f(t)$ given by (8), we obtain

$$f'(t)=p(A_1)+2p(A_2)t+3p(A_3)t^2+\cdots+np(A_n)t^{n-1}. \tag{11}$$

Formulas (10) and (11) together lead to:

$$p(A_1)+2p(A_2)t+3p(A_3)t^2+\cdots+np(A_n)t^{n-1}=np(pt+q)^{n-1}. \tag{12}$$

Set $t=1$ into both sides of (12). Since $p+q=1$, we obtain the equality (6).

Now multiply both sides of (12) by t. We find

$$p(A_1)t+2p(A_2)t^2+3p(A_3)t^3+\cdots+np(A_n)t^n=np(pt+q)^{n-1}t. \tag{13}$$

Let us find the derivatives of both sides of (13). The derivative of the left-hand side is found by means of formula (15) of Chapter II. We obtain the polynomial

$$p(A_1)+2^2p(A_2)t+\cdots+n^2p(A_n)t^{n-1}.$$

The derivative of the right-hand side can be evaluated by the rule d) for the derivative of a product from Section 2, Chapter II. Write the right-hand side of (13) in the form of a product: $(np(tp+q)^{n-1})\cdot t$. By rule d) the derivative of this expression is $(np(tp+q)^{n-1})'\cdot t+(np(tp+q)^{n-1})\cdot t'$. By formula (15) of Chapter II, we have $t'=1$; by rule c) of Section 2 of Chapter II we have $(np(tp+q)^{n-1})'=np((tp+q)^{n-1})'$ and by formula (19) of Chapter II we have $((tp+q)^{n-1})'=(n-1)(tp+q)^{n-2}p$ since $(tp+q)'=p$, by formula (15) of Chapter II. Therefore, equating the derivatives of the left and right-hand sides of (13) we obtain

$$p(A_1)+2^2p(A_2)t+\cdots+n^2p(A_n)t^{n-1}=np(pt+q)^{n-1}+n(n-1)p^2(tp+q)^{n-2}. \tag{14}$$

Set $t = 1$ into (14). On the left we get σ_2. On the right (in view of $p + q = 1$) we get $np + n(n-1)p^2 = n^2p^2 + np(1-p) = n^2p^2 + npq$ (since $1 - p = q$).

We can now turn to the proof of Chebyshev's theorem. Chebyshev's device was to write inequality (2), which defines the necessary indices, in the form

$$\left|\frac{k - np}{\varepsilon n}\right| > 1,$$

i.e.

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

and then to multiply each term $p(A_k)$ in the sum S_ε by $\left(\frac{k - pn}{\varepsilon n}\right)^2$, which is greater than 1, and therefore increases the sum. After that he considered the *total* sum $\overline{S}_\varepsilon$ of *all* terms $\left(\frac{k - pn}{\varepsilon n}\right)^2 p(A_k)$, $k = 0, 1, \ldots, n$, and not only those for indices k which satisfy (2). It is clear that the sum $\overline{S}_\varepsilon$ differs from the sum S_ε by a certain number of positive terms, and so it must be greater than S_ε.

Hence, $S_\varepsilon < \overline{S}_\varepsilon$. Now, by quite elementary transformations (using the Lemma) we can evaluate the sum $\overline{S}_\varepsilon$ exactly, and thus we obtain the wanted inequality for the sum S_ε.

Therefore, we have to find the sum $\overline{S}_\varepsilon$ of all terms $\left(\frac{k - pn}{\varepsilon n}\right)^2 p(A_k)$ for $k = 0, 1, \ldots, n$. Their common denominator $(\varepsilon n)^2$ can be taken out and the expressions $(k - pn)^2$ can be expanded: $(k - pn)^2 = k^2 - 2npk + p^2n^2$. Every term in the sum $\overline{S}_\varepsilon$ (after $(\varepsilon n)^2$ has been taken out) gives three terms. The sum of the first terms is σ_2 on the left of (7). The sum of the second terms, after the common factor $-2pn$ has been taken out, is the sum σ_1 defined by (6). Finally, the sum of the third terms, after p^2n^2 has been taken out is σ_0, defined by (5). Adding up all the obtained equalities, we find the expression for the sum $\overline{S}_\varepsilon$:

$$\overline{S}_\varepsilon = \frac{1}{\varepsilon^2 n^2}(\sigma_2 - 2pn\sigma_1 + p^2n^2\sigma_0).$$

Substituting the values obtained for σ_2, σ_1 and σ_0 in the Lemma, we find

$$\overline{S}_\varepsilon = \frac{1}{\varepsilon^2 n^2}(n^2p^2 + npq - 2p^2n^2 + p^2n^2) = \frac{pq}{\varepsilon^2 n}. \tag{15}$$

As we saw, $S_\varepsilon < \overline{S}_\varepsilon$ and therefore $S_\varepsilon < \frac{pq}{\varepsilon^2 n}$ and the proof of Chebyshev's inequality is finished.

Let us briefly analyse the method which lies in the essence of this proof. The sum S_ε which we want to estimate has a perfectly simple form. The difficulty lies in the fact that the sum is formed by terms which are chosen according to a rather strange criterion (the indices k have to satisfy (2)). The first thing that comes to mind is to ignore these conditions and to take the sum of *all* terms. This sum is easily evaluated: according to the Lemma, it is equal to 1. But it is too large and does not lead to the equality we want. Chebyshev's device was to introduce the additional factor $\left(\frac{k-np}{\varepsilon n}\right)^2$ and only *after* that to consider the sum of all terms, ignoring the restriction (2). In this process the terms which appear in the sum S_ε are increased, but those which do not are decreased so much that the total sum $\overline{S}_\varepsilon$ becomes sufficiently small (namely, for the terms which do not appear in S_ε we have $\left(\frac{k-np}{\varepsilon n}\right)^2 < 1$).

We have met here with a phenomenon which is very often present in mathematics. Namely, important and interesting inequalities usually follow from an *identity* after an obvious estimate. This obvious estimate in our case is the inequality $S_\varepsilon \leqslant \overline{S}_\varepsilon$ and the identity is the relation (15) which gives the explicit expression for the sum $\overline{S}_\varepsilon$. This is how inequalities of fundamental importance in mathematics are proved. But sometimes they are proved in a different way—this might indicate that there is an underlying identity which we do not yet know.

Return once more to the formulation of Chebyshev's theorem. As we already explained, we are considering the event that in Bernoulli's scheme I^n the event a occurs k times, where either $k > np + n\varepsilon$, or $k < np - n\varepsilon$; in other words, we do not consider the event that in Bernoulli's scheme the event a occurs k times where $np - n\varepsilon \leqslant k \leqslant np + n\varepsilon$. We found that the first event has small probability (for large n), not exceeding $pq/\varepsilon^2 n$. This means that the second event has greater probability, not less than $1 - (pq/\varepsilon^2 n)$. For example, consider a series of large number of repetitions of one experiment under constant conditions. Suppose that one experiment can have only two outcomes—a and b, where the probability of a is p. This situation (if the number of experiments is n) is described, as we saw, by Bernoulli's scheme (I^n, p). The experiment may be, for instance testing a large set of objects (animals, technical details, etc) for a given property, knowing that p-th part of the set has this property. The scheme I^n describes the possible results of the testing. According to Chebyshev's theorem, in the series of n experiments the number of occurrences of the outcome a will be between $np - n\varepsilon$ and $np + n\varepsilon$ with a probability greater than $1 - \frac{p(1-p)}{\varepsilon^2 n}$. Here, ε can be any number which we can choose as we like. For example, let $p = \frac{3}{4}$. Choosing $\varepsilon = \frac{1}{100}$, we see that in the series of n experiments the number k of occurrences of the event a will satisfy the inequality $\frac{3}{4}n - \frac{n}{100} \leqslant k \leqslant \frac{3}{4}n + \frac{n}{100}$ with the probability not less than

$$1 - \frac{\frac{3}{4}\cdot\frac{1}{4}}{\left(\frac{1}{100}\right)^2 n}.$$

Since $\frac{3}{4^2} < \frac{2}{10}$, this probability is not less than

$$1 - \frac{\frac{2}{10}}{\left(\frac{1}{100}\right)^2 n} = 1 - \frac{2000}{n}.$$

For $n = 200\,000$, this probability will be not less than 0,99. The number of occurrences of the event a after $200\,000$ experiments which have this large probability will be between 148 000 and 152 000 (since $\dfrac{3}{4}n = 150\,000$, $n \cdot \dfrac{1}{100} = 2000$, $np - n\varepsilon = 148\,000$, $np + n\varepsilon = 152\,000$).

Conversely, using Chebyshev's theorem we can estimate the number of experiments to be made in order to obtain the probability p accurately enough. Suppose that we want to determine it with accuracy up to 1/10 and that the probability it is equal to the obtained number is not less than 0,99. According to Chebyshev's theorem we have to put $\varepsilon = 1/10$ and to use the inequality

$$\frac{pq}{\left(\frac{1}{10}\right)^2 \cdot n} < 0{,}01.$$

Notice that $q = 1-p$, and for any p such that $0 \leqslant p \leqslant 1$, we have $pq = p(1-p) \leqslant 1/4$. This follows from the fact that the geometric mean is not greater than the arithmetic mean of the numbers p, q, which is 1/2. Therefore, it is enough that n should satisfy the inequality

$$\frac{\frac{1}{4}}{\left(\frac{1}{10}\right)^2 \cdot n} < 0{,}01$$

which implies $n > 2500$.

Problems

1. In the set of some objects, 95% of them have a certain property. Prove that among 200 000 objects, the number of those which have this property is between 189 000 and 191 000 with probability not less than 0,99.

2. Modify Problem 1 so that the portion of objects which have a certain property is not known. What is the probability that after testing 100 objects we can determine it with accuracy up to 0,1?

3. For any positive integer $r \leqslant n$ find the sum of all terms

$$k(k-1)\cdots(k-r+1)p(A_k)$$

for $k = 1, \ldots, n$.

4. For $r \leqslant 4$ evaluate the sums σ_r consisting of terms $k^r p(A_k)$ for all $k = 0, 1, 2, 3, 4$. Do this in two different ways: a) by the reasoning of the proof of the Lemma, and b) by expressing the sums σ_r in terms of sums evaluated in Problem 3 for $r = 1, 2, 3, 4$.

5. Try to improve the inequality (4) in Chebyshev's theorem, applying the factor $\left(\dfrac{k-np}{n\varepsilon}\right)^4$ instead of $\left(\dfrac{k-np}{n\varepsilon}\right)^2$. The improvement will be that n^2 will appear in the denominator of the right-hand side of the inequality instead of n.

I. R. Shafarevich,
Russian Academy of Sciences,
Moscow, Russia

THE TEACHING OF MATHEMATICS
2000, Vol. III, 2, pp. 63–82

SELECTED CHAPTERS FROM ALGEBRA

I. R. Shafarevich

Abstract. This paper is the fourth part of the publication "Selected chapters from algebra", the first three having been published in previous issues of the Teaching of Mathematics, Vol. I (1998), 1–22, Vol. II, 1 (1999), 1–30, Vol. II, 2 (1999), 65–80, and Vol. III, 1 (2000), 15–40.

AMS Subject Classification: 00 A 35

Key words and phrases: Primes, function $\pi(n)$, Chebyshev inequality, asymptotical law of distribution of primes.

CHAPTER IV. PRIMES

1. Infinity of the number of primes

In this chapter we return to the question which we have already dealt with in Chapter I. It was proved there that each natural number can be uniquely represented as a product of primes. Therefore, when the multiplication is concerned, the primes are the simplest elements and all the natural numbers can be obtained by multiplying primes, similarly to the fact that they can be obtained by the operation of addition starting from the number 1. From this point of view, the interest for the set of all primes can be easily understood. There are four primes among the first ten natural numbers: 2, 3, 5, 7. Further primes can be found by dividing each of the consequent numbers by previously found primes, in order to decide whether it is a prime itself. In this way we find the following 25 primes among the first one hundred natural numbers:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97.$$

How far does this sequence continue? The question arose already in the antique times. The answer was given by Euclid:

Theorem 1. *The number of primes is infinite.*

We give several proofs of this theorem.

First proof—the one contained in Euclid's "Elements". Suppose we have found n primes: $p_1, p_2, \dots, p_n$. Consider the number $N = p_1 p_2 \cdots p_n + 1$. As we saw

This paper is an English translation of: И. Р. Шафаревич, *Избраные главы алгебры. Глава IV. Простые числа*, Математическое образование, N 1(4), янв.–мар. 1998, Москва, стр. 2–21. In the opinion of the editors, the paper merits wider circulation and we are thankful to the author for his kind permission to let us make this version.

in §2 of Chapter I, each number has at least one prime divisor. In particular, N has a prime divisor. But none of the numbers $p_1, p_2, \dots, p_n$ can divide N. To see this, let p_i be a divisor of N. Then $N - p_1 \cdots p_n$ must be divisible by p_i, but since $N - p_1 \cdots p_n = 1$, this is impossible. It follows that this prime divisor must be different from each p_i, $i = 1, \dots, n$, which means that after each n primes there must be at least one additional prime. This proves the theorem.

Second proof. According to the theorem of the section "Set algebra" of Chapter III, the number of numbers which are smaller than the given number N and relatively prime with it, is given by the formula

$$N\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\cdots\left(1-\frac{1}{p_n}\right), \tag{1}$$

where $p_1, \dots, p_n$ are all prime divisors of N. We shall prove the theorem by contradiction. Suppose that the number of primes is finite and that $p_1, \dots, p_n$ are all of them. Set $N = p_1 \cdots p_n$. Substituting in formula (1) we obtain for each factor $p_i(1 - \frac{1}{p_i})$ the expression $p_i - 1$, and for the whole product (1) the expression $(p_1 - 1)(p_2 - 1) \cdots (p_n - 1)$. As we know that there exist primes greater than 2 (for example, 3), the number obtained is *greater* than 1. Hence, there exists a number a, smaller than N, relatively prime with N and different from 1. But a has at least one prime divisor which must be contained among the numbers $p_1, \dots, p_n$, and so a cannot be relatively prime with N. We obtained a contradiction which proves the theorem.

The infinite sequence of primes is, on the other hand, very sparsely distributed among natural numbers. For example, there are arbitrary big "gaps" in this sequence, i.e., one can find (successively further away) any given number of consecutive numbers which are not prime. For example, n numbers $(n+1)!+2$, $(n+1)!+3$, $\dots$, $(n+1)!+n+1$ are obviously not primes—the first one is divisible by 2, the second by 3, the last by $n+1$.

For some time mathematicians have searched for a formula expressing primes. For example, Euler found an interesting polynomial $x^2 + x + 41$, which, for 40 values of x—from 0 to 39—obtains prime values. However, it is obvious that for $x = 40$ its value is a nonprime number 41^2. It is not hard to conclude that there cannot exist a polynomial $f(x)$ which takes prime values for all natural values $x = 0, 1, 2, \dots$ (not even speaking about the possibility that its values are *all* of the primes). We shall show this on an example of a polynomial of second degree $ax^2 + bx + c$ with integer coefficients a, b, c. Suppose that for $x = 0$ the polynomial has a prime value c. Then for each $x = kc$ its value $ak^2c^2 + bkc + c$ is divisible by c. This value can be equal to c for at most one additional value of k (besides $k = 0$), which can be easily checked. Moreover, there does not exist a polynomial $f(x) = ax^2 + bx + c$ having prime values for each integer x, *starting from some limit.* Indeed, suppose that the values of the polynomial $f(x)$ are prime for each $x \geqslant m$. Set $x = y + m$, $f(y+m) = g(y)$; then all the values of the polynomial $g(y)$ are prime for all integers $y \geqslant 0$, by the assumption, and its coefficients are also integers, since $g(y) = a(y+m)^2 + b(y+m) + c$. The same reasoning also applies to a polynomial of an arbitrary degree n: $f(x) = a_0 + a_1x + \cdots + a_nx^n$. If all of its values for integer

$x \geqslant 0$ are prime, it means that $f(0) = a_0 = p$ is prime, too. Then for each integer k the values $f(kp) = p + a_1 kp + \cdots + a_n(kp)^n$ are divisible by p. They can be equal to p only if $p + a_1 kp + \cdots + a_n(kp)^n = p$, i.e., $a_1 + a_2 kp + \cdots + a_n(kp)^{n-1} = 0$, and the last equation in k is of the degree $n-1$, and according to Theorem 3 of Chapter II it has at most $n-1$ roots. For all other values of k the number $f(kp)$ is divisible by p and different from p, i.e., it is not a prime.

If we suppose that the values of the polynomial $f(x)$ are prime only for integer values of $x \geqslant m$, for a certain number m, then we can set $x = y+m$ and $f(y+m) = g(y)$. The polynomial $g(y) = a_0 + a_1(y+m) + \cdots + a_n(y+m)^n$ is obtained by expanding all the parentheses by the binomial formula and reducing similar terms. Therefore its coefficients are again integers, but it obtains prime values for *all* integers $y \geqslant 0$, which again is a contradiction.

It can be also proved that for an arbitrary number k no polynomial in k variables with integer coefficients exists such that all of its values for all natural values of its variables are primes. Nevertheless, it appears that there is a polynomial of degree 25 with 26 variables, having the following property: if we select those values of that polynomial which are obtained for nonnegative integer values of its variables and which are positive themselves, then the set of such values coincides with the set of primes. Since 26 is equal to the number of letters of the Latin alphabet, it is possible to denote the variables by the letters: a, b, $\ldots$, x, y, z. Then the polynomial is of the form:

$$\begin{gathered}
F(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z) = \\
= (k+2)\{1 - [wz+h+j-q]^2 - [(gk+2g+k+1)(h+j)+hz]^2 - \\
-[2n+p+q+z-e]^2 - [16(k+1)^3(k+2)(n+1)^2+1-f^2]^2 - \\
-[e^3(t+2)(a+1)^2+1-o^2]^2 - [(a^2-1)y^2+1-x^2]^2 - \\
-[16r^2y^4(a^2-1)+1-u^2]^2 - [(a+u^2(u^2-a^2)-1)(n+4dy)^2+1-(x+cu)^2]^2 - \\
-[n+l+v-y]^2 - [(a-1)l^2+1-m^2]^2 - [ai+k+1-l-i]^2 - \\
-[p+l(a-n-1)+b(2an+2a-n^2-2n-2)-m]^2 - \\
-[q+y(a-p-1)+s(2ap+2a-p^2-2p-2)-x]^2 - \\
-[z+pl(a-p)+t(2ap-p^2-1)-pm]^2\}.
\end{gathered}$$

This polynomial has been written here just to impress the reader. Its number of variables is too big. It can be proved that it takes also negative values $-m$, where m is not prime. Hence, it does not give us information about the sequence of primes either.

Long trials convinced the majority of mathematicians that there is no easy formula describing the sequence of primes. There exist "explicit formulae" describing the primes, but they use objects which are even less known than the primes themselves. That is why mathematicians concentrated on the characteristics of the sequence of primes "in total" and not "in parts". We will deal with this kind of questions in the next sections.

Problems

1. Prove that there are infinitely many primes of the form $3s+2$.

2. The same for the primes of the form $4s+3$.

3. Prove that each two numbers $2^{2^n}+1$ and $2^{2^m}+1$ are relatively prime. Deduce once more the infinity of the number of primes. [*Hint.* Assuming that p is a common divisor of two such numbers, find the remainders of division of 2^{2^m} and 2^{2^n} by p.]

4. Let $f(x)$ be a polynomial with integer coefficients. Prove that there exist infinitely many distinct prime divisors of its values $f(1)$, $f(2)$, $\dots$. (If you do not succeed immediately, solve the problem for the polynomials of the first and of the second degree.)

5. Denote by p_n the n-th prime in the natural order. Prove that $p_{n+1} < p_n^n + 1$.

6. Using the notation of Problem 5, prove that $p_n < 2^{2^n}$. Deduce the similar inequality $p_n \leqslant 2^{2^n}+1$ from the result of Problem 3.

7. Using the notation of Problem 5, prove that $p_{n+1} < p_1 p_2 \cdots p_n$.

2. Euler's proof of the infinity of the number of primes

We shall give another proof of the infinity of the number of primes, which is due to Euler, and which clarifies some general properties of this sequence.

Let us start with the "prehistory", that is, with some simple facts which had been known before Euler started dealing with questions about primes. The question is about how big the following sums can be:

$$1, \quad 1+\frac{1}{2}, \quad 1+\frac{1}{2}+\frac{1}{3}, \quad \dots, \quad 1+\frac{1}{2}+\dots+\frac{1}{n}, \quad \dots$$

Using notation from section 3 of Chapter II, these are the sums $(Sa)_n$, where a is the sequence of the inverses of natural numbers $1, \frac{1}{2}, \frac{1}{3}, \dots$. Since we denoted the sums of the m-th powers of natural numbers from 1 to $n-1$ by $S_m(n)$ (cf. formula (28) of Chapter II), it is natural to denote our sums by $S_{-1}(n)$.

We have come to a concept which we shall often deal with later, so we consider it now in more detail. It refers generally to properties of an *infinite* sequence of positive numbers $s_1, s_2, \dots, s_n, \dots$ (in our case it appeared as the sequence of sums of another sequence, but for the moment that is of no importance). One type of such sequences is called *bounded.* This means that there exists a number C (the same for the whole sequence), such that $s_n < C$ for all $n = 1, 2, 3, \dots$. If the sequence does not have this property, it is called *unbounded.* This means that *no* number C can possess this property, i.e., for each number C there exists an index n such that $s_n \geqslant C$. Finally, it may happen that for each number C there exists an index n such that $s_m \geqslant C$ for all $m = n, n+1, \dots$. In other words, for n sufficiently large, the numbers s_n become arbitrary large. In that case the sequence is called *unboundedly increasing.* For example, the sequence 1, 2, 1, 3, 1, 4, … , where 1 stands on odd places, and natural numbers stand on even places in succession, is unbounded, but not unboundedly increasing, since one can find the number 1 arbitrarily far in it.

If a sequence $a = a_1, a_2, \dots, a_n, \dots$ of positive numbers is given, and $s = Sa$, then $s_{n+1} > s_n$ (since $s_{n+1} = s_n + a_{n+1}$, $a_{n+1} > 0$), and, generally, $s_m > s_n$ for $m > n$. Therefore, such a sequence will be unboundedly increasing if it is unbounded. For example, if all $a_i = 1$, then $s_n = n$ and the sequence $s_1, s_2, \dots$ is unbounded. But in other cases it may be bounded. An example can be visualised on Fig. 1, where we first divide the segment between 0 and 1 in half and set $a_1 = \frac{1}{2}$, then divide again the segment between 0 and $\frac{1}{2}$ in half and set $a_2 = \frac{1}{4}$, etc. In this way, $a_n = \frac{1}{2^n}$. The result of adding such numbers is represented on Fig. 1 and it is obvious that the sums S_n stay inside our initial segment, i.e., $S_n < 1$.

Fig. 1

It is easy to check the last assertion by calculation. If $a_n = \frac{1}{2^n}$, then

$$(Sa)_n = \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n} = \frac{1}{2}\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{n-1}}\right),$$

and by formula (12) of Chapter I

$$(Sa)_n = \frac{1}{2}\,\frac{\frac{1}{2^n} - 1}{\frac{1}{2} - 1} = 1 - \frac{1}{2^n},$$

so that $(Sa)_n < 1$ for each n.

We shall show now that in the case of the sequence $1, \frac{1}{2}, \frac{1}{3}, \dots$ the *first* case appears: although the terms of the sequence decrease, they do not decrease fast enough, and their sums (i.e., $S_{-1}(n)$) increase unboundedly.

Lemma 1. *The sum $S_{-1}(n)$ is, for n sufficiently large, greater than an arbitrary given number.*

Let the number k be given. We assert that for some n (and so also for all greater integers) $S_{-1}(n) > k$. Take n such that $n - 1 = 2^m$ for some m. Divide the sum

$$S_{-1}(n) = 1 + \left(\frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \cdots + \left(\frac{1}{2^{m-1}+1} + \cdots + \frac{1}{2^m}\right)$$

in parts as it is shown: in groups contained between two consecutive powers of two. Each parenthesis has the form $\frac{1}{2^{k-1}+1} + \cdots + \frac{1}{2^k}$, and the number of parentheses is equal to m. In each parenthesis we replace each summand by the smallest one entering that parenthesis, that is by the last one. Since the number of summands in such a parenthesis is equal to $2^k - 2^{k-1} = 2^{k-1}$, we obtain that the k-th parenthesis is greater than $\frac{2^{k-1}}{2^k} = \frac{1}{2}$. As a result, we obtain that $S_{-1}(n) > 1 + \frac{m}{2}$. This inequality is valid for each n if $n-1 = 2^m$. It remains to put $1 + \frac{m}{2} = k$, i.e., $m = 2k-1$ and $n = 2^{2k-1}+1$. Then $S_{-1}(n) > k$.

Now we come to Euler's proof. His idea is connected with the method of computing the sums of powers of the divisors of a natural number, which was described in section 3 of Chapter I (cf. formula (13) in Chapter I). Denote the sum of k-th powers of all divisors (including 1 and n) of a natural number n by $\sigma_k(n)$. According to formula (13) of Chapter I, for the number n having canonical factorisation $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$,

$$\sigma_k(n) = \frac{p_1^{k(\alpha_1+1)}-1}{p_1^k-1}\frac{p_2^{k(\alpha_2+1)}-1}{p_2^k-1}\cdots\frac{p_r^{k(\alpha_r+1)}-1}{p_r^k-1}. \tag{2}$$

Formula (2) had been known since the antique times, but it was implicitly assumed that the number k in it was positive. Finally, Euler got interested in it and he posed the question—what would happen if k was integer, but negative? The answer is, of course, that there is no difference, the derivation of formula (2) is completely formal and the same for negative as well as for positive values of k. In particular, it is valid for $k=-1$. The sum of (-1)-st powers (i.e., the inverses) of the divisors of a given number n will be denoted, as before, by $\sigma_{-1}(n)$. Formula (2) gives

$$\sigma_{-1}(n) = \frac{1-\dfrac{1}{p_1^{\alpha_1+1}}}{1-\dfrac{1}{p_1}}\cdot\ldots\cdot\frac{1-\dfrac{1}{p_r^{\alpha_r+1}}}{1-\dfrac{1}{p_r}}$$

(we interchanged the order of summands in numerators and denominators in each of the fraction). From here (since all the expressions in numerators are less than 1),

$$\sigma_{-1}(n) < \frac{1}{\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\cdots\left(1-\frac{1}{p_r}\right)}. \tag{3}$$

Let us now replace n in this formula by $n!$ ($p_1, \ldots, p_r$ are now prime divisors of $n!$). The numbers $1, 2, \ldots, n$ are all contained among the divisors of $n!$. Therefore, the sum $\sigma_{-1}(n!)$ definitely contains summands $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$, whose sum is equal to $S_{-1}(n+1)$. According to Lemma 1, already the sum $S_{-1}(n+1)$ is greater than any given number k for n sufficiently large. Since other summands in the sum $\sigma_{-1}(n!)$ are positive, the same conclusion is valid for it. If the number of primes were finite and $p_1, \ldots, p_r$ were the whole list of them, we would obtain that

$$\frac{1}{\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\cdots\left(1-\frac{1}{p_r}\right)} > k,$$

where k is an arbitrary number. This is, of course, a contradiction.

The value of the above proof is not that the assumption of finiteness of the number of primes has led to a contradiction, but that it, when the infinity of that number has already been proved, gives some quantitative characteristics of the sequence of primes. Namely, reformulating the result obtained, we can now say that if $p_1, p_2, \ldots, p_n, \ldots$ is the infinite sequence of primes, then the expression $\frac{1}{\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\cdots\left(1-\frac{1}{p_n}\right)}$ becomes greater than any arbitrary number for n sufficiently large. This is, of course, equivalent to the fact that the denominator of the last fraction becomes *smaller* than an arbitrary positive number for n sufficiently large. We have proved

Theorem 2. *If $p_1, p_2, \ldots, p_n, \ldots$ is the sequence of all primes, then the product $\left(1-\dfrac{1}{p_1}\right)\left(1-\dfrac{1}{p_2}\right)\cdots\left(1-\dfrac{1}{p_n}\right)$, for n sufficiently large, becomes smaller than any given positive number.*

This is a first approximation to our goal. Let us try now to give a more useful form of the characteristic obtained.

Theorem 3. *If $p_1, p_2, \ldots, p_n, \ldots$ is the sequence of all primes, then the sequence of sums $\dfrac{1}{p_1}+\dfrac{1}{p_2}+\cdots+\dfrac{1}{p_n}$ increases unboundedly.*

Derivation of Theorem 3 from Theorem 2 is purely formal: it does not use the fact that $p_1, p_2, \ldots, p_n, \ldots$ is the sequence of *primes*—it could be an arbitrary sequence of natural numbers which satisfies the conditions of Theorem 2.

Lemma 2. *For each natural number $n > 1$ the inequality*

$$1-\frac{1}{n} \geqslant \frac{1}{4^{1/n}} \tag{4}$$

is valid.

Since both sides of inequality (4) are positive, rasing them to the power of n, we obtain an *equivalent* inequality

$$\left(1-\frac{1}{n}\right)^n \geqslant \frac{1}{4}, \tag{5}$$

which we are going to prove. Expanding the left-hand side by the binomial formula we obtain

$$\left(1-\frac{1}{n}\right)^n = 1 - n\,\frac{1}{n} + \frac{n(n-1)}{2}\,\frac{1}{n^2} - \frac{n(n-1)(n-2)}{3!}\,\frac{1}{n^3} + \cdots + (-1)^n\frac{1}{n^n}. \tag{6}$$

Absolute values of the summands on the right-hand side of formula (6) form the sequence $C_n^k\frac{1}{n^k}$. We examined such a sequence in connection with the Bernoulli scheme in the section "Language of probability" in Chapter III (formula (35)). More precisely, if in that formula we put $p = \dfrac{1}{n+1}$, $q = 1 - \dfrac{1}{n+1} = \dfrac{n}{n+1}$, then we obtain that $p+q = 1$, $p^k q^{n-k} = (n+1)^{-n} n^{n-k}$ and the numbers obtained differ from the ones examined in formula (6) just by the common factor $\left(\frac{n}{n+1}\right)^n$. The expression $(n+1)p-1$ is in our case equal to zero. In the section "Language of probability" of Chapter III we proved that if $k > (n+1)p-1$ (in our case $k > 0$), then the $(k+1)$-st term is smaller than the k-th one. This means that the numbers of the sequence $C_n^k \frac{1}{n^k}$, $k = 1, 2, \ldots, n$ decrease monotonously. (We referred here to Chapter III just to stress the connection between different problems that we are dealing with. It would, of course, be easy to write down the ratio of the $(k+1)$-st term of the sequence to the k-th one and conclude that it is less than 1). We can see that in formula (6), the first two terms on the right-hand side cancel. The next two terms (after cancellation which can be done easily) give $\frac{1}{3} - \frac{1}{3n^2}$. This number is not less than $\frac{1}{4}$ for $n \geqslant 2$ (check it yourself!). The rest of the terms can be grouped in pairs, where in each pair the first term is positive and the next negative, but, as we have seen, by absolute value less than the first one. Therefore each pair gives a positive contribution to the sum (6). If n is odd, then the number of summands on the right-hand side of formula (6) is even (it is equal to $n+1$) and the sum is partitioned into $\frac{n+1}{2}$ pairs. If n is even, then after grouping into pairs, there remains the summand $\frac{1}{n^n}$. In such a way, in any case the right-hand side consists of a summand which is not less than $\frac{1}{4}$, and some additional positive summands. This proves inequality (5), and so the lemma itself.

Theorem 3 is now evident. For each p_i we have, according to the Lemma:

$$1 - \frac{1}{p_i} \geqslant \frac{1}{4^{1/p_i}}.$$

Multiplying these inequalities for $i = 1, \ldots, n$ we obtain

$$\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\cdots\left(1 - \frac{1}{p_n}\right) \geqslant \frac{1}{4^{\left(\frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_n}\right)}}.$$

If the sums $\frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_n}$ were for each n less than a certain value k, it would follow that

$$\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\cdots\left(1 - \frac{1}{p_n}\right) \geqslant \frac{1}{4^k}.$$

This contradicts Theorem 2.

We run here into a problem of a new kind. If N is a subset of a finite set M, then we can tell how much "smaller" N is than M, comparing the number of their elements, e.g., computing the ratio $n(N)/n(M)$. But now we have two infinite sets: the set of all natural numbers and the set of all primes contained in it. How can we compare them? Theorem 3 offers one way of comparing, not very easy at first sight. It can be applied to each sequence of natural numbers a: $a_1, a_2, \ldots, a_n, \ldots$. According to Lemma 1, for the sequence of all natural numbers, the sums of their inverses (i.e., the sums $S_{-1}(n)$) increase unboundedly. We can think of the sequence a to be "tightly" distributed among natural numbers if it has the same property, i.e., if the sums $\frac{1}{a_1}, \frac{1}{a_1} + \frac{1}{a_2}, \ldots, \frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}, \ldots$ unboundedly increase. This means that in the sequence a enough natural numbers remained so that the sums of their inverses are not too much less than the sums $S_{-1}(n)$ of the inverses of all natural numbers. If, on the other hand, the sums of inverses of the sequence a remain bounded, we can think of it as “loosely” distributed in the natural row. Theorem 3 states that the sequence of primes is “tight”. The most “loose” case is the case of a sequence a having only a finite number of terms.

But there are intermediate cases. For instance, the sequence of squares: 1, 4, 9, $\dots$, n^2, $\dots$. It is natural to denote the corresponding sums $1+\frac{1}{4}+\frac{1}{9}+\dots+\frac{1}{n^2}$ by $S_{-2}(n)$. We shall prove that they are bounded by a number not depending on n. We use the same idea as in the proof of Lemma 1. Let m be such that $2^m \geqslant n$. Then $S_{-2}(n) \leqslant S_{-2}(2^m)$. We divide the sum $S_{-2}(2^m) = 1+\frac{1}{2^2}+\frac{1}{3^2}+\dots+\frac{1}{2^{2m}}$ into parts:

$$(1)+\left(\frac{1}{2^2}\right)+\left(\frac{1}{3^2}+\frac{1}{4^2}\right)+\dots+\left(\frac{1}{(2^{m-1}+1)^2}+\dots+\frac{1}{2^{2m}}\right).$$

Each part $\dfrac{1}{(2^{k-1}+1)^2}+\dots+\dfrac{1}{2^{2k}}$ again contains 2^{k-1} terms and the first term is the greatest. Therefore this part cannot be greater than $2^{k-1}\dfrac{1}{(2^{k-1}+1)^2} < 2^{k-1}\dfrac{1}{(2^{k-1})^2} = \dfrac{1}{2^{k-1}}$. Therefore, $S_{-2}(2^m) \leqslant 1+1+\dfrac{1}{2}+\dfrac{1}{2^2}+\dots+\dfrac{1}{2^{m-1}} = 1+\dfrac{1-\frac{1}{2^m}}{1-\frac{1}{2}} \leqslant 1+\dfrac{1}{1-\frac{1}{2}} = 3$. So, none of the sums $S_{-2}(n)$ is greater than 3.

In such a way, Theorem 3 shows that, for example, the primes are distributed more “tightly” in the natural row than the squares.

Problems

1. Prove that for each given integer $k \geqslant 2$ and for all natural n, the sums $S_{-k}(n) = \frac{1}{1^k}+\frac{1}{2^k}+\dots+\frac{1}{n^k}$ are bounded.

2. Let the sequence a be an arithmetic progression: $a_0 = p$, $a_1 = p+q$, $a_2 = p+2q$, $\dots$, $a_n = p+nq$ for some natural p and q. Prove that the sums $\frac{1}{a_0}$, $\frac{1}{a_0}+\frac{1}{a_1}$, $\dots$, $\frac{1}{a_0}+\frac{1}{a_1}+\dots+\frac{1}{a_n}$, $\dots$ become unboundedly large for n sufficiently large.

3. Let the sequence a be a geometric progression: $a_0 = c$, $a_1 = cq$, $a_2 = cq^2$, $\dots$, $a_n = cq^n$, $\dots$, where c and q are natural numbers. Is it “tight” or “loose” in the natural row?

4. Let $p_1, \dots , p_n, \dots$ be the sequence of all primes. Prove that the expressions $\dfrac{1}{\left(1-\frac{1}{p_1^2}\right)\left(1-\frac{1}{p_2^2}\right)\cdots\left(1-\frac{1}{p_n^2}\right)}$ are bounded for each n.

3. The function $\pi(n)$

In this section we will try once more to estimate how much the sequence of primes differs from the sequence of all natural numbers. We will replace the more elaborate method of comparing "tight" and "loose" sequences from the previous section by a more naive one, which can be understood more easily. Namely, we will try to answer the naive question—"which portion of the sequence of natural numbers is covered by the primes"— by finding how many primes there are smaller than 10, how many smaller than 100, how many smaller than 1000, etc. For each natural number n, denote by $\pi(n)$ the number of primes not greater than n, so that $\pi(1) = 0$, $\pi(2) = 1$, $\pi(4) = 2, \ldots$. What can be said about the ratio $\frac{\pi(n)}{n}$ when n increases?

First of all, consider what can be learned from tables. Each assertion or question concerning natural numbers can be checked for all natural numbers not exceeding a certain limit N. This fact plays a role in the number theory, which investigates properties of natural numbers, similar to that of the possibility of experimenting in theoretical physics. in particular, one can compute the values $\pi(n)$ for $n = 10^k$, $k = 1, 2, \ldots, 10$. The following table is obtained.

n	$\pi(n)$	$\dfrac{n}{\pi(n)}$
10	4	2.5
100	25	4.0
1000	168	6.0
10000	1229	8.1
100000	9592	10.4
1000000	78498	12.7
10000000	664579	15.0
100000000	5761455	17.4
1000000000	50847534	19.7
10000000000	455059512	22.0

Table 1.

We see that the ratio $\frac{n}{\pi(n)}$ is constantly increasing, which means that $\frac{\pi(n)}{n}$ is decreasing all the time. In other words, the portion of primes among the first n numbers becomes close to zero when n increases. According to the tables, it could be said that "the primes constitute a zero portion among all natural numbers". That was the way Euler formulated this fact, although his reasoning did not contain a full proof. We will now give the precise formulation and then the proof.

Theorem 4. *The ratio $\frac{\pi(n)}{n}$ becomes smaller than any given positive number for n sufficiently large.*

In order to prove the theorem we have to estimate somehow the function $\pi(n)$. For actual calculation of its values we start with the prime 2, then we cancel all the numbers which are multiples of 2 and not exceeding n. Then we take the first remaining number—this will be 3—and repeat the process. We continue till we have exhausted all the numbers not exceeding n. The numbers which are not cancelled (2, 3, etc.) are all primes not exceeding n. This method was used already in the antique times; it is called "the sieve of Eratosthenes".

We will apply the same process in our reasoning. Suppose we have already found the first r primes: p_1, p_2, ..., p_r. Then the remaining primes, not exceeding n, are contained among "noncancelled" numbers, not exceeding n, i.e., among those numbers $m \leqslant n$ which are not divisible by any of the numbers p_1, p_2, ..., p_r. But the number of numbers not exceeding n and not divisible by any of the primes p_1, p_2, ..., p_r was explored in Chapter III—it is given by the formula in the section Set algebra of Chapter III. As we showed there, the expression in the formula can be replaced with the easier one $n\left(1-\frac{1}{p_1}\right)\cdots\left(1-\frac{1}{p_r}\right)$, where the error is less than 2^r (formula (28) of Chapter III). Hence, the number s of numbers $m \leqslant n$ not divisible by any of the primes p_1, p_2, ..., p_r satisfies the inequality

$$s \leqslant n\left(1-\frac{1}{p_1}\right)\cdots\left(1-\frac{1}{p_r}\right)+2^r. \tag{7}$$

All $\pi(n)$ primes not exceeding n are contained either among r primes $p_1, p_2, \ldots, p_r$, or among s numbers accounted for by inequality (7). In such a way, $\pi(n) \leqslant s+r$, and

$$\pi(n) \leqslant n\left(1-\frac{1}{p_1}\right)\cdots\left(1-\frac{1}{p_r}\right)+2^r+r. \tag{8}$$

Inequality (8) is remarkable because it contains the product $\left(1-\frac{1}{p_1}\right)\cdots\left(1-\frac{1}{p_r}\right)$ which can be estimated using Theorem 2.

Now we can pass to the proof of Theorem 4. Let an arbitrary small positive number ε be given. We have to find a number N, depending on ε, such that $\frac{\pi(n)}{n} < \varepsilon$ is valid for each $n > N$. In the inequality (8) we replace r by a greater number 2^r (cf. Problem 6 in section 2 of Chapter I), in order to obtain a simpler inequality

$$\pi(n) \leqslant n\left(1-\frac{1}{p_1}\right)\cdots\left(1-\frac{1}{p_r}\right)+2^{r+1}. \tag{9}$$

In the inequality (9) there are two summands and we shall choose N so that for each $n \geqslant N$ each of the summands will not exceed $\varepsilon n/2$. Then from the inequality (9) we will conclude that $\pi(n) < \varepsilon n$, and so $\frac{\pi(n)}{n} < \varepsilon$. Recall that till now the number r in our reasoning was arbitrary. We choose it so that the first summand does not exceed $\varepsilon n/2$, and then we choose N such that the second summand does not exceed $\varepsilon n/2$. The first choice is possible according to Theorem 2. It states that for r sufficiently large, the product $\left(1-\frac{1}{p_1}\right)\cdots\left(1-\frac{1}{p_r}\right)$ is less than any arbitrary given positive number. We can take $\varepsilon/2$ to be such a number. Then the first summand in the inequality (9) does not exceed $\varepsilon n/2$. The second summand can be dealt with even more easily. Now r has already been chosen. Choose N such that $2^{r+1} < eN/2$. For this it is enough to choose $N > \dfrac{2^{r+2}}{\varepsilon}$. Then $2^{r+1} < \dfrac{\varepsilon N}{2} \leqslant \dfrac{\varepsilon n}{2}$ for each $n \geqslant N$. Theorem 4 is proved.

Note that if we choose an arithmetic progression $am+b$ even with a very big difference a, i.e., being very "sparse", then the number of terms of this progression not exceeding n is the same as the number of integers m satisfying $am \leqslant n-b$, i.e., $\left[\frac{n-b}{a}\right]$. We saw in section 3 of Chapter III that $\left[\frac{n-b}{a}\right]$ differs from $\frac{n-b}{a}$ by not more than 1. Hence, the number of terms of the progression not exceeding n is not less than $\frac{n-b}{a}-1$. Its quotient with n is not less than $\frac{1}{n}(\frac{n-b}{a}-1) = \frac{1}{a} - \frac{1}{n}\frac{b}{a} - \frac{1}{n}$. When n increases, this number approaches $\frac{1}{a}$ and is not becoming arbitrarily small. Thus, Theorem 4 would become false if we replaced the sequence of primes in it by an arbitrary arithmetic progression. This shows that primes are distributed more sparsely than any arithmetic progression.

Problems

1. Let p_n denote the n-th prime. Prove that for an arbitrarily large positive number C the inequality $p_n > Cn$ is valid for n sufficiently large. [*Hint.* Use the fact that $\pi(p_n) = n$.]

2. Consider natural numbers having the property that, when written in decimal form, they do not contain a certain digit (e.g., 0). Let $q_1, q_2, \ldots, q_n, \ldots$ be those numbers, written in ascending order, and let $\pi_1(n)$ denotes the number of such numbers not exceeding n. Prove that the ratio $\frac{\pi_1(n)}{n}$ becomes smaller than any arbitrary given positive number, for n sufficiently large. Prove that the sums $\frac{1}{q_1}, \frac{1}{q_1}+\frac{1}{q_2}, \ldots, \frac{1}{q_1}+\frac{1}{q_2}+\cdots+\frac{1}{q_n}, \ldots$ are bounded. [*Hint.* Do not try to copy the proof of Theorem 4. Split the sum to parts, where in each part denominators are contained between 10^k and 10^{k+1}. Then find the number of numbers q_i in such intervals. The answer depends on the digit which is excluded: $r=0$ or $r \neq 0$.]

APPENDIX

Inequalities of Chebyshev for $\pi(n)$

We have put this text in the Appendix mainly for formal reasons, because we have to use logarithms, the knowledge of which is not assumed in the rest of the text. Recall that the *logarithm with basis a of the number x* is a number y such that

$$a^y = x.$$

This is written as $y = \log_a x$. In the sequel it will always be assumed that $a > 1$ and that x is a positive number. Basic properties of logarithms, following directly from the definition, are:

$$\log_a(xy) = \log_a x + \log_a y, \quad \log_a c^n = n \log_a c, \quad \log_a a = 1.$$

$\log_a x > 0$ if and only if $x > 1$. Logarithm is a monotonous function, i.e., $\log_a x \leqslant \log_a y$ if and only if $x \leqslant y$.

In this text, if the basis of a logarithm is not indicated, it is supposed that it is equal to 2; $\log x$ means $\log_2 x$.

The second reason for putting this text in the Appendix is the following. In the rest of the book, the logic of reasoning was clear, namely, why do we follow a certain road (at least I hope it was so). Here we encounter the case, not rare in mathematical investigations, when it looks as if some new consideration has "jumped in from nowhere", when even the author cannot explain how he came to the conclusion. About such situations Euler used to say: "Sometimes it seems to me that my pencil is smarter than myself". Of course, these are results of long trials and unknown work of psyche.

We are going now to continue our investigations concerning the ratio $\frac{\pi(n)}{n}$ when n is increasing unboundedly. Take another look at Table 1, showing the values of $\pi(n)$ for $n = 10^k$, $k = 1, 2, \dots, 10$. The last column of the table contains the values of the ratio $\frac{n}{\pi(n)}$ for some values of n. We see that when we pass from $n = 10^k$ to $n = 10^{k+1}$, that is when we go down by one row, the value of $\frac{n}{\pi(n)}$ increases always approximately by the same value. Namely, the first number is equal to 2.5; the second differs from it for 1.5, and the further differences are: 2; 2.1; 2.3; 2.3; 2.3; 2.4; 2.3; 2.3. We see that all of these numbers are close to the one: 2.3. Not trying at the moment to explain the meaning of this particular value, let us suppose that also beyond the range of our table the numbers $\frac{n}{\pi(n)}$, when passing from $n = 10^k$ to $n = 10^{k+1}$, increase by amounts which are closer and closer to a certain constant α. This would mean that $\frac{n}{\pi(n)}$ for $n = 10^k$ would be very close to αk. But, if $n = 10^k$, then by the definition $k = \log_{10} n$. It is natural to assume that also for other values of n the ratio $\frac{n}{\pi(n)}$ is very close to $\alpha \log_{10} n$. Thus, $\pi(n)$ is very close to $c\,\dfrac{n}{\log_{10} n}$, where $c = \alpha^{-1}$.

A lot of mathematicians where attracted by the secret of distribution of primes and tried to solve it using tables. In particular, Gauss got interested in this question almost as a child. His interest in mathematics started, it seems, from the child's interest in numbers and forming tables. In general, a lot of great mathematicians showed virtuosity in calculations and were capable of doing immense ones, often by heart (Euler was even fighting insomnia in that way!). When Gauss was only 14, he constructed a table of primes (in fact, a smaller one than our Table 1) and came to the conclusion we have formulated. Later on this conclusion was considered by many mathematicians. But the first result in that direction was proved only half a century later, in 1850, by Chebyshev.

Chebyshev proved the following assertion.

Theorem. *There exist constants c and C such that for all $n > 1$*

$$c\,\frac{n}{\log n} \leqslant \pi(n) \leqslant C\,\frac{n}{\log n}. \tag{10}$$

Before we proceed with the proof, we give some remarks concerning the formulation of the theorem. What is the basis of the logarithms we are using? The answer is: arbitrary. From the definition of logarithms it immediately follows that $\log_b x = \log_b a \log_a x$: it is enough to substitute a by $b^{\log_b a}$ in the relation

$a^{\log_a x} = x$, to obtain $b^{\log_b a \log_a x} = x$, which shows that $\log_b x = \log_b a \log_a x$. Hence, if the inequality (10) is proved for $\log_a n$, then it is true also for $\log_b n$, with the substitution of c and C by $\frac{c}{\log_b a}$ and $\frac{C}{\log_b a}$.

Inequalities (10) express the idea inspired by tables that $\pi(n)$ is "close" to $c\,\frac{n}{\log n}$ for some c. The question why in our hypothetical reasoning there appeared one constant c, and in the theorem there appear two of them—c and C—and whether it is possible to use only one constant, will be discussed after the proof of the theorem.

The key to the proof of Chebyshev's theorem are properties of binomial coefficients C_n^k: mostly the fact that they are integers and some properties about their divisibility by primes. We shall recall these properties before we pass to the proof.

First of all, there is a proposition proved in section 3 of Chapter II which says that the sum of all binomial coefficients C_n^k for $k = 0, 1, \ldots, n$ is equal to 2^n. Since the sum of positive summands is greater than any of them, we deduce that

$$C_n^k \leqslant 2^n. \tag{11}$$

We shall particularly need large binomial coefficients. We saw in Chapter II that for even $n = 2m$ the coefficient C_{2m}^m is greater than all the others, and for odd $n = 2m + 1$ there exist two equal coefficients C_{2m+1}^m and C_{2m+1}^{m+1} which are greater than the others. We draw our attention to them, particularly to

$$C_{2n}^n = \frac{2n(2n-1)\cdots(n+1)}{1 \cdot 2 \cdot \ldots \cdot n}. \tag{12}$$

If we group the factors of the numerator with the factors of the denominator taken in reverse order, we obtain

$$C_{2n}^n = \frac{2n}{n} \cdot \frac{2n-1}{n-1} \cdot \ldots \cdot \frac{n+1}{1}.$$

Obviously, no factor in the last formula is less than 2, so

$$C_{2n}^n \geqslant 2^n. \tag{13}$$

Consider now properties of divisibility of binomial coefficients by primes. Factors in the numerator in the expression (12) are obviously divisible by all primes greater than n and not exceeding $2n$. These primes cannot divide any factor in the denominator, and so they do not cancel and they are divisors of C_{2n}^n. The number of primes between n and $2n$ is equal to $\pi(2n) - \pi(n)$ and all of them are greater than n, hence

$$C_{2n}^n \geqslant n^{\pi(2n)-\pi(n)}. \tag{14}$$

An analogous assertion is valid for the "middle" coefficients $C_{2n+1}^n = C_{2n+1}^{n+1}$ with an odd lower index. If we write them as

$$C_{2n+1}^n = \frac{(2n+1)\cdots(n+2)}{1 \cdot 2 \cdot \ldots \cdot n},$$

we see that $\pi(2n+1)-\pi(n+1)$ of primes, greater than $n+1$ and not exceeding $2n+1$, enters the numerator and cannot be cancelled with the denominator. Since they are greater than $n+1$, we have

$$C_{2n+1}^{n} > (n+1)^{\pi(2n+1)-\pi(n+1)}. \tag{15}$$

The inequalities (14) and (15) already reveal important connections between binomial coefficients and prime numbers.

Finally, we state the last of the properties of binomial coefficients which we need for the proof; although it is quite easy, it is not as obvious as the previous ones.

Lemma. *For an arbitrary binomial coefficient C_n^k, any power of a prime dividing it does not exceed n.*

We draw the attention to the fact that we are not speaking about the *exponent* of the power but about the *power itself.* In other words, we assert that if p^r divides C_n^k, where p is a prime, then $p^r \leqslant n$. For example, $C_9^2 = 9 \cdot 4$ is divisible by 9 and by 4, and both of these numbers do not exceed 9.

Write the binomial coefficient in the form

$$C_n^k = \frac{n(n-1)\dots(n-k+1)}{1\cdot 2\cdot \ldots \cdot k}. \tag{16}$$

The prime p we are dealing with has to divide the numerator of this fraction. Denote by m the factor in numerator which contains the maximal power of p (or one of those having such a property), and by p^r this maximal power. Obviously, $n \geqslant m \geqslant n-k+1$. Set $n-m=a$, $m-(n-k+1)=b$, then $a+b=k-1$ and C_n^k can be written in the form

$$C_n^k = \frac{(m+a)(m+a-1)\cdots(m+1)m(m-1)\cdots(m-b)}{k!}. \tag{17}$$

The factor m is now the most important for us and we write down the product in the numerator as having a factors to the left and b factors to the right of it. Rearrange the denominator analogously: $k! = (1\cdot 2\cdot \ldots \cdot a)(a+1)\cdots(a+b)(a+b+1)$. Since $(a+1)(a+2)\cdots(a+b)$ is divisible by $b!$, this product (denominator) has the form $a!\,b!\,l$, where l is an integer. Now we can rewrite C_n^k in the following form

$$C_n^k = \frac{m+a}{a}\cdot\frac{m+a-1}{a-1}\cdot\ldots\cdot\frac{m+1}{1}\cdot\frac{m-1}{1}\cdot\ldots\cdot\frac{m-b}{b}\cdot\frac{m}{l}, \tag{18}$$

where we transferred the factor $\frac{m}{l}$ to the end.

Note that in each of the factors $\frac{m+i}{i}$ or $\frac{m-j}{j}$ $(i=1,\dots,a,\ j=1,\dots,b)$ the power of p entering the numerator completely cancels with the denominator, and hence after cancellation only the denominator could be divisible by p (though it can also be relatively prime with p). Really, consider, for instance, fractions $\frac{m+i}{i}$ (factors of the type $\frac{m-j}{j}$ can be treated in the same way). Let i be divisible exactly by p^s, i.e., $i=p^s u$, where u is relatively prime with p. If $s<r$, then $m+i$ is

also divisible exactly by p^s: setting $m = p^r v$ (recall that m is divisible by p^r), we obtain that $m + i = p^s(u + p^{r-s}v)$. If $s \geqslant r$, then for the same reasons $m + i$ is divisible by p^r and taking into account the way m was chosen (it is divisible by the greatest power of p of all the numbers between $n - k + 1$ and n and this power is p^r), we conclude that the number $m + i$ cannot be divisible by a greater power of p than the r-th. Thus, p^r cancels and in the numerator there remains a number not divisible by p. As a result we see that among all the factors in the expression (18), only the last one can contain p as a factor. But the power of p which divides m is p^r, and this means the product (18) cannot be divisible by a greater power of p than p^r. Since p^r divides m, and $m \leqslant n$, it is $p^r \leqslant n$. The Lemma is proved.

Let us see what it says about the canonical factorisation $C_n^k = p_1^{\alpha_1} \cdots p_m^{\alpha_m}$. First of all, the primes $p_1, \ldots, p_m$ can appear just from the numerator of the expression (16), therefore all $p_i \leqslant n$ and so $m \leqslant \pi(n)$. According to the Lemma, $p_i^{\alpha_i} \leqslant n$ for $i = 1, \ldots, m$. As a result we obtain that

$$C_n^k \leqslant n^{\pi(n)}. \tag{19}$$

Now we can proceed with the proof of the Chebyshev's theorem itself, i.e., with the proof of the inequalities (10). Note that it is enough to prove these inequalities for all values of n starting with a certain fixed limit n_0. For all $n < n_0$ these inequalities can then be obtained by decreasing the constant c and increasing the constant C. If we wanted to obtain values of these constants explicitly and in the most economic way, then we could check, using tables of primes, that the inequalities (10) are valid for values $n \leqslant n_0$ (in our arguments n_0 will not be a large number).

We start with juxtaposition of inequalities (13) and (19) for the binomial coefficient C_{2n}^n. We obtain that $2^n \leqslant C_{2n}^n \leqslant (2n)^{\pi(2n)}$ and hence

$$2^n \leqslant (2n)^{\pi(2n)}. \tag{20}$$

Taking logarithms with basis 2 of both sides (recall that we will write $\log_2 x = \log x$) and using monotonicity of the logarithm, we obtain that $n \leqslant \pi(2n) \log 2n$ and so

$$\pi(2n) \geqslant \frac{n}{\log 2n} = \frac{1}{2}\frac{2n}{\log 2n},$$

i.e., the left of the two inequalities (10) with the constant $c = \frac{1}{2}$. But for the time being it is proved only for even values of n. For odd values of the form $2n + 1$ we use the monotonicity of the logarithm and of the function $\pi(n)$. It follows that

$$\pi(2n+1)\log(2n+1) \geqslant \pi(2n)\log 2n.$$

Substituting here the obtained inequality for $\pi(2n)$, we see that

$$\pi(2n+1) \geqslant \frac{n \log 2n}{\log(2n)\log(2n+1)} = \frac{n}{\log(2n+1)}.$$

Since it is always $n \geqslant \frac{1}{3}(2n+1)$, it follows that

$$\pi(2n+1) \geqslant \frac{1}{3}\frac{2n+1}{\log(2n+1)}.$$

Thus, the left inequality (10) is proved for odd n with the constant $c = \frac{1}{3}$. So, the left inequality (10) is valid for all n and $c = \frac{1}{3}$.

We proceed to the proof of the right inequality in (10). We shall prove it by induction on n. Let, first of all, n be even. We will write $2n$ instead of it. Taking the inequality (11) for the coefficient C_{2n}^n (i.e., substitute in C_n^k n by $2n$ and k by n) together with the inequality (14), as a consequence we obtain

$$n^{\pi(2n)-\pi(n)} \leqslant 2^{2n}$$

and, passing to logarithms,

$$\pi(2n) - \pi(n) \leqslant \frac{2n}{\log n}, \qquad \pi(2n) \leqslant \pi(n) + \frac{2n}{\log n}. \tag{21}$$

In accordance with the inductive hypothesis, suppose that our inequality has been proved: $\pi(n) \leqslant C\frac{n}{\log n}$ with a constant C whose value we shall make more precise later. Substituting in the formula (21), we obtain:

$$\pi(2n) \leqslant C\,\frac{n}{\log n} + \frac{2n}{\log n} = \frac{(C+2)n}{\log n}.$$

We would like to prove the inequality $\pi(2n) \leqslant \frac{C\cdot 2n}{\log 2n}$ and for that we have to choose the constant C in such a way that the inequality

$$\frac{(C+2)n}{\log n} \leqslant \frac{2Cn}{\log 2n} \tag{22}$$

is valid for all n, starting from some limit.

This is just a simple school exercise. Cancel in the inequality both sides by n, remark that $\log 2n = \log 2 + \log n = \log n + 1$ and denote $\log n$ by x. Then the inequality (22) takes the form

$$\frac{C+2}{x} \leqslant \frac{2C}{x+1}.$$

Multiplying both sides by $x(x+1)$ (as $x > 0$) and transforming, we write it in the form $(C-2)x \geqslant C+2$. Obviously, C has to be chosen so that $C - 2 > 0$. Setting, e.g., $C = 3$, we obtain that it is valid for $C = 3$ and all $x \geqslant 5$. Since x denotes $\log n$, this means that the necessary inequality would be valid if $n \geqslant 2^5 = 32$, $2n \geqslant 64$.

It remains to consider the case of odd values of the form $2n+1$. Compare the inequality (11) (substituting in it n by $2n+1$ and k by n) with the inequality (15). We obtain the inequality

$$2^{2n+1} \geqslant (n+1)^{\pi(2n+1)-\pi(n+1)}$$

and, taking logarithms, the inequality

$$2n+1 \geqslant (\pi(2n+1) - \pi(n+1))\log(n+1).$$

From here, using the inductive hypothesis about $\pi(n+1)$, we obtain, as before

$$\pi(2n+1) \leqslant C\,\frac{n+1}{\log(n+1)} + \frac{2n+1}{\log(n+1)}.$$

The inequality that we need: $\pi(2n+1) \leqslant C\frac{2n+1}{\log(2n+1)}$ will be proved if we can check that

$$C\,\frac{n+1}{\log(n+1)} + \frac{2n+1}{\log(n+1)} \leqslant C\,\frac{2n+1}{\log(2n+1)} \tag{23}$$

for a suitable choice of the constant C and for all n starting from some limit. This is again an exercise of purely school type, though a bit harder than the previous one. In order to compare various terms in the inequality more easily, replace on the left-hand side $2n+1$ by a greater value $2(n+1)$:

$$C\,\frac{n+1}{\log(n+1)} + \frac{2n+1}{\log(n+1)} \leqslant \frac{(C+2)(n+1)}{\log(n+1)}. \tag{24}$$

In order to transform the right-hand side, note that $2n+1 \geqslant \frac{3}{2}(n+1)$ for $n \geqslant 1$, $\log(2n+1) \leqslant \log(2n+2) = \log(n+1)+1$. Hence,

$$\frac{2n+1}{\log(2n+1)} \geqslant \frac{(3/2)(n+1)}{\log(n+1)+1}. \tag{25}$$

Comparing inequalities (24) and (25) we see that the inequality (23) will be proved if we prove that

$$\frac{(C+2)(n+1)}{\log(n+1)} \leqslant \frac{(3/2)C(n+1)}{\log(n+1)+1}.$$

Cancelling both sides by $n+1$ and putting $\log(n+1) = x$, we arrive at the inequality

$$\frac{C+2}{x} \leqslant \frac{(3/2)C}{x+1},$$

which can be solved completely in the same way as in the previous case. It is enough to multiply both sides by $x(x+1)$ and reduce similar terms. We obtain the inequality $(C+2)x + C + 2 \leqslant \frac{3}{2}Cx$, i.e., $(\frac{1}{2}C - 2)x \geqslant C + 2$. Setting $C = 6$, we see that the inequality is valid for $x \geqslant 8$, i.e., for $n+1 \geqslant 2^8$, $2n+1 \geqslant 511$. Thus, the right inequality (10) is proved with the constant $C = 6$ and for all values of n starting with 511. The Theorem is proved.

Note that Theorem 4 appears as an easy consequence of the Theorem just proved. Really, since $\pi(n) < C\frac{n}{\log n}$, we have $\frac{\pi(n)}{n} \leqslant \frac{C}{\log n}$. And as a logarithm changes monotonously and increases unboundedly ($\log 2^k = k$), $\frac{\pi(n)}{n}$ becomes less than any arbitrary positive number. But, the proof of the theorem of Chebyshev was based on completely different considerations than the proof of Theorem 4.

At the end, we return once more to assertions which can be made by considering Table I. Starting from it, we came to the claim that $\frac{n}{\pi(n)}$ is close to $\log_{10} n$ with a certain value of the constant C: the first decimal figures of the number C^{-1} are

2.3. Hence it can be concluded that $\pi(n)$ is close to $C^{-1}\frac{n}{\log_{10} n}$. This expression can be given a simpler form $\frac{n}{\log_e n}$, if we use a new basis of logarithms e such that $C\log_{10} n = \log_e n$. But, as it was said earlier, it is always $\log_b x = \log_b a \cdot \log_a x$, and so our relation will be fulfilled if $C = \log_e 10$. Substituting the value $x = b$ into the relation $\log_b x = \log_b a \cdot \log_a x$, we obtain that $\log_b a \cdot \log_a b = 1$ and the relation $C = \log_e 10$ which we are interested in can be rewritten as $C^{-1} = \log_{10} e$.

14-year-old Gauss turned his attention to these relations and tried to guess which number e could be, so that $\log_{10} e$ is close to $(2.3)^{-1}$. Such a number at that time was well known, thanks to the fact that the logarithm with such a basis has a lot of useful properties. This number is commonly denoted by e. The logarithm with the basis e is called *natural* and is denoted by ln: $\log_e x = \ln x$. Here, to the end of this page, we have to consider that the reader is familiar with the concept of the natural logarithm.

In such a way, a natural assertion which can be deduced from tables is that $\pi(n)$ becomes close to $\frac{n}{\ln n}$ when n increases unboundedly. The theorem of Chebyshev which has been just proved states (if we use natural logarithms) that there exist two such constants c and C, that $c\frac{n}{\ln n} < \pi(n) < C\frac{n}{\ln n}$, starting from some n. The hypothetical refinement deduced from tables asserts that the inequalities $c\frac{n}{\ln n} < \pi(n) < C\frac{n}{\ln n}$ are valid for n large enough *whichever constants $c < 1$ and $C > 1$ we take.* This assertion is called *the asymptotical law of distribution of primes.* It was stated by Gauss and some other mathematicians at the end of XVIII and the beginning of XIX century. After the proof of the inequalities of Chebyshev in 1850 it seemed that all that was needed was a better choice and approaching of the constants c and C. However, the asymptotical law of distribution of primes was proved just half a century later, at the end of XIX century, using completely new ideas, proposed by Riemann.

Problems

1. Prove that $p_n > an\log n$ for a certain constant $a > 0$ [*Hint.* Use the fact that $\pi(p_n) = n$.]

2. Prove that $\log n < \sqrt{n}$, starting from some limit (find it). [*Hint.* Reduce the problem to proving the inequality $2^x > x^2$ for real x, starting from some limit. Let $n \leqslant x \leqslant n+1$, where n is an integer. Reduce to proving the inequality $2^n \geqslant (n+1)^2$ and use the induction.]

3. Prove that $p_n < Cn^2$ for some constant C. [*Hint.* Apply the inequality of the previous problem and use the fact that $n = \pi(p_n)$.]

4. Prove that $p_n < An\log n$ for some constant A.

5. Prove that that the largest exponent a for which p^a divides $n!$ is equal to $\left[\frac{n}{p}\right] + \left[\frac{n}{p^2}\right] + \dots + \left[\frac{n}{p^k}\right]$. Here $\left[\frac{r}{s}\right]$ is the incomplete quotient of dividing r by s, the sum extends to all k for which $p^k \leqslant n$, p denotes an arbitrary prime and n an arbitrary natural number.

6. Using the result of Problem 5 give a new proof of the Lemma in the Appendix.

7. Prove that if $p_1, \ldots, p_r$ are all the primes between m and $2m+1$, then their product does not exceed 2^{2m}.

8. Determine the constants c and C such that the inequality (10) is valid for all n.

9. Try to find as large as possible a constant c and as small as possible a constant C, for which the inequality (10) is valid for all n, starting from some limit.

I. R. Shafarevich,
Russian Academy of Sciences,
Moscow, Russia

THE TEACHING OF MATHEMATICS
2001, Vol. IV, 1, pp. 1–34

SELECTED CHAPTERS FROM ALGEBRA

I. R. Shafarevich

Abstract. This paper is the fifth part of the publication “Selected chapters from algebra”, the first four having been published in previous issues of the Teaching of Mathematics, Vol. I (1998), 1–22, Vol. II, 1 (1999), 1–30, Vol. II, 2 (1999), 65–80, Vol. III, 1 (2000), 15–40 and Vol. III, 2 (2000), 63–82.

AMS Subject Classification: 00 A 35

Key words and phrases: Real numbers, limits, infinite sums, decimal representations, real roots of polynomials, Sturm's theorem.

CHAPTER V. REAL NUMBERS AND POLYNOMIALS

1. Axioms of real numbers

In the present chapter we shall try to make our idea of real numbers more precise. Our tendency will not be towards very rigorous reasoning, but we shall only try to give enough accuracy to our notions and reasoning in this field, so that we are able to *prove* statements about real numbers.

If we choose an origin and a unit on a line, we can represent real numbers as points on the line. Thus, if we make our idea of real numbers more precise, we give at the same time a more precise description of a line and points lying on it. In the sequel we shall often, as an illustration, use this bijective correspondence between real numbers and points on a line.

Let us try to take geometry as an example and bring the precision of definitions and arguing to the level which already exists in the school geometry courses. There, some axioms appear as the basis of all the construction, and starting from these axioms all other statements are proved. Axioms themselves are not proved: we take them on the basis of experiment or intuition.

In order to be more concrete, let us look at the construction of *plane geometry* based on axioms. We can distinguish three types of logical notions. First of all, there are basic geometrical notions—points and lines. Then, there are basic relations: a point lies on a line; a point lies on a line between other two points. Neither of these are defined. We think as if a “list” of all points and all lines exists

This paper is an English translation of: И. Р. Шафаревич, *Избраные главы алгебры. Глава V. Действительные числа и многочлены*, Математическое образование, N 2(5), апр.–июнь 1998, Москва, стр. 31–66. In the opinion of the editors, the paper merits wider circulation and we are thankful to the author for his kind permission to let us make this version.

somewhere, and we know which points lie on which lines or which triples of points A, B, C on the line l are such that B lies between A and C. And only in third place there are axioms, i.e., statements about basic notions and relations among them. For instance: each two distinct points belong to exactly one line. Or: among three distinct points on a line, there is exactly one lying between other two.

There is a complete analogy with real numbers. The basic notions here are real numbers themselves. This means that, for the moment, we do not assume anything more about real numbers, but only that they constitute a certain set. Basic relations between real numbers are of two different types: operations and inequalities. Let us describe them in more detail.

1) Operations with real numbers

For every two real numbers a and b we define a third number c, called the *sum* of a and b. We write this as: $a + b = c$.

For every two real numbers a and b we define a third number d, called the *product* of a and b. We write this as: $ab = d$.

2) Inequalities between real numbers

For some pairs of real numbers a and b we have that a is less than b. We write this as: $a < b$. The same relation is also written as $b > a$. If we want to say that $a < b$ or $a = b$, we write $a \leqslant b$ (or $b \geqslant a$).

Before we pass to the formulation of axioms connecting basic notions with basic relations among them, let us emphasize once more the analogy with geometry. Write analogous notions in the table:

Algebra		Geometry
	Basic notions	
Real numbers		Point, line, ...
	Basic relations	
sum: $a + b = c$		A point lies on a line.
product: $ab = d$		Point C lies between
inequality: $a < b$		points A and B.
		...
	Axioms	
...		...

There is no need to list geometrical axioms here; and axioms on real numbers shall be listed now. They will be formulated in terms of basic notions and relations between them, listed in the table. We group the axioms according to the basic relations they deal with.

I (axioms of addition)

I$_1$. Commutative law: $a + b = b + a$ for arbitrary real numbers a and b.

I$_2$. Associative law: $a + (b + c) = (a + b) + c$ for arbitrary real numbers a, b and c.

I$_3$. There exists a number called *zero*, denoted by 0, such that $a + 0 = a$ is valid for each real number a.

(Remark. There exists exactly one such number. If $0'$ were another number with the same property, we would have $0' + 0 = 0'$, by the definition of 0, $0' + 0 = 0 + 0'$ by the commutative law and $0 + 0' = 0$, by the definition of $0'$. Finally, we obtain $0' = 0' + 0 = 0 + 0' = 0$, i.e., $0' = 0$.)

I_4. For each real number a there exists a number called opposite, denoted by $-a$, such that $a + (-a) = 0$.

(Remark. For the given number a there exists exactly one such number. If a' were another number with the same property: $a + a' = 0$, we would have $(a + (-a)) + a' = 0 + a' = a'$. Also, $(a + (-a)) + a' = ((-a) + a) + a'$, and by the associative law, $((-a) + a) + a' = (-a) + (a + a')$. By the property of number a', $a + a' = 0$ and $(-a) + 0 = -a$. Taking these equalities together, we obtain that $a' = -a$.)

II (axioms of multiplication)

II_1. Commutative law: $ab = ba$ for arbitrary real numbers a and b.

II_2. Associative law: $a(bc) = (ab)c$ for arbitrary real numbers a, b and c.

II_3. There exists a number called *unit*, denoted by 1, such that $a \cdot 1 = a$ for an arbitrary real number a.

(Remark. There exists only one such number. It can be proved in the same way as the remark following axiom I_3—we only have to replace addition by multiplication, and 0 by 1.)

II_4. For each real number a, different from 0, there exists a number called inverse, denoted by a^{-1}, such that $a \cdot a^{-1} = 1$.

(Remark. For each real number a different from 0, there exists only one such number. The proof is exactly the same as in the remark following axiom I_4.)

III (axiom of addition and multiplication)

III_1. Distributive law: $(a + b)c = ac + bc$ for arbitrary real numbers a, b and c.

IV (axioms of order)

IV_1. For any two real numbers a and b exactly one of the following three relations holds: $a = b$ or $a < b$ or $b < a$.

IV_2. If for some three real numbers a, b and c we have $a < b$ and $b < c$, then $a < c$.

IV_3. If $a < b$, then $a + c < b + c$ for arbitrary three real numbers a, b and c.

IV_4. If $a < b$ and $c > 0$, then $ac < bc$ for arbitrary three real numbers a, b and c.

V (real and rational numbers)

Rational numbers are contained among real numbers, and operations and inequalities, defined for real numbers, when applied to rational ones, give usual operations and inequalities.

VI (axiom of Archimedes)

For each real number a there exists a natural number n such that $a < n$.

VII (axiom of embedded segments)

> Let a_0, a_1, a_2, ... and b_0, b_1, b_2, ... be two sequences of real numbers, satisfying $a_0 \leqslant a_1 \leqslant a_2 \leqslant \cdots$, $b_0 \geqslant b_1 \geqslant b_2 \geqslant \cdots$ and $b_n \geqslant a_n$ for each n. Then there exists a real number c, such that $b_m \geqslant c$ and $c \geqslant a_n$ for all m and n.

If we use representation of real numbers on a line, then numbers x satisfying the condition $a \leqslant x$ and $x \leqslant b$ ($a \leqslant x \leqslant b$ for short) are represented by the set which is called a *segment* and denoted by $[a, b]$. So, the premises of the last axiom state that the segments $I_n = [a_n, b_n]$ are embedded one into another: $I_0 \supset I_1 \supset I_2 \supset \cdots$. The axiom states that there exists a point (i.e., a number) which is common for all these embedded segments (hence the name of the axiom).

All the usual properties of real numbers easily follow from the listed axioms. It would be too boring to devote several pages to these completely obvious arguments. Hence, we shall only formulate some assertions which we shall need later—and give just some remarks in connection with their proofs (see also problems 2, 3, 4).

It follows from the axioms of group II that for each number a different from 0 and each number b, the number $c = a^{-1}b$ is the unique solution of the equation $ax = b$. It is called the *quotient* of b and a and denoted by $\frac{b}{a}$. All the usual rules about dealing with parentheses and fractions follow from the axioms.

Since for a natural number n the equality $n = 1 + \cdots + 1$ (n summands) is valid, it follows from the axioms of group III that for each number a, the number na (product of n and a) is equal to the sum $a + \cdots + a$ (n summands).

Axiom IV$_3$ implies that if $a < b$ and $c < d$, then $a + c < a + d < b + d$. If $a < 0$, then $-a > 0$ (because from $-a < 0$ it would follow $0 < 0$). As a result we conclude that each real number is either positive ($a > 0$), has the form $-b$, where $b > 0$, when we say that it is negative, or it is equal to 0. Multiplication obeys the usual "rule of signs". As usual, we write $|x| = x$ if $x \geqslant 0$ and $|x| = -x$ when $x < 0$.

Axiom of embedded segments (axiom VII) is particularly useful when the length of segment I_n (i.e., the difference $b_n - a_n$) becomes arbitrary small when n increases. In other words, if for an arbitrary real number $\varepsilon > 0$ there exists an index N such that $b_n - a_n < \varepsilon$ for all $n \geqslant N$. In such a case one can conclude more than just what is said in the axiom:

Lemma 1. *If differences $b_n - a_n$ become arbitrary small with increasing of the index n, then number c, whose existence is guaranteed by axiom VII, is unique.*

Proof. Suppose that there exist two such numbers: c and c' and, for example, $c < c'$. Then $a_n < c < c' < b_n$ and $c' - c = b_n - a_n - (c - a_n) - (b_n - c') \leqslant b_n - a_n$. We obtain (for n sufficiently large) that $c' - c < \varepsilon$ for an arbitrary given number $\varepsilon > 0$. For instance, such a relation has to be valid for $\varepsilon = \frac{c' - c}{2}$, whence $\frac{1}{2}(c' - c) < 0$, but this contradicts the fact that $c' - c > 0$, $\frac{1}{2} > 0$.

We meet exactly this situation when we intend to measure the given real number approximately, with deficiency or excess, using rational numbers. In that case a_n and b_n are rational numbers. An example is the construction of $\sqrt{2}$ we spoke about in Section 1 of Chapter I. Thus, axiom VII formulates what we intuitively have in mind when we speak about "better and better measuring". Together with the preceding Lemma it gives us the possibility of *constructing* real numbers with the prescribed properties. We shall often use this observation later.

Concerning axioms V and VI we just remark that we assume here natural and, more generally, rational numbers to be known. We shall not analyse these notions in detail.

Let us remark at the end that the given axioms are not *independent.* This means that some of them could be proven as theorems, relying on other axioms (see, e.g., problem 6). We have just gathered those properties of real numbers which we are used to and which are intuitively convincing. Taking greater number of axioms we obtained the right to skip not very interesting proofs of some intuitively obvious facts.

Problems

1. Which of the axioms I–VII are also valid in the set of rational numbers, and which are specific for real numbers?

2. Prove, using axioms I–III, that for each real number a, $0a = 0$.

3. Prove that for arbitrary real numbers a and b the equation $a + x = b$ has a solution and that it is unique.

4. Prove that for arbitrary real numbers $a \neq 0$ and b the equation $ax = b$ has a solution and that it is unique.

5. Consider the set of rational numbers as a subset of the set of real numbers—on the basis of axiom V. Prove that rational number 0 coincides with the real number 0 whose existence is based on axiom I_3. Do the same for rational number 1 and the real number 1 whose existence is based on axiom II_3.

6. Not using axiom V, prove that numbers 0, 1, $1 + 1$, $\dots$, $1 + 1 + \dots + 1$ (n summands) are different for all natural n. Here 1 denotes the number whose existence is guaranteed by axiom II_3. Hence, prove that natural numbers are contained amongst the reals, and that operations and inequalities, defined for real numbers, when applied to natural ones, give usual operations and inequalities. Prove after that the assertion of axiom V. In that way, this axiom is in fact superfluous in our list, since it could be proven on the basis of other axioms.

7. Instead of the operation of multiplication, given by definition for real numbers, define a new operation $\odot$ given by the formula $a \odot b = a + b + ab$. Does it obey the axioms of group II?

2. Limits and infinite sums

In order to illustrate the role of axiom of embedded segments as a method of construction of new real numbers, we shall introduce several notions which will also be useful later.

We met in Chapter IV sequences which were bounded as well as sequences which increased unboundedly. Consider now sequences which are decreasing. For the sake of simplicity, consider first sequences of positive numbers and call such a sequence *unboundedly decreasing* if its terms unboundedly approach zero. The exact definition can be made analogously to the definition of unboundedly increasing sequences, given in Section 2, Chapter IV.

A sequence a_n of nonnegative real numbers is said to *approach zero unboundedly* if for each arbitrary small positive number ε there exists a natural number N such that $a_n < \varepsilon$ for all $n > N$. In such a case we also say that the sequence a_n *tends* to 0 and denote it by: $a_n \to 0$ when $n \to \infty$ ("when n tends to infinity").

A typical example of such a sequence is the sequence $a_n = \frac{1}{n}$.

Consider now a less obvious example.

Lemma 2. *If a is an arbitrary positive number smaller than 1, then the sequence $a_n = a^n$ unboundedly approaches* 0, *i.e., $a^n \to 0$ when $n \to \infty$.*

Really, put $a = 1/A$. Then $A > 1$ and it can be written in the form $A = 1 + x$ with $x > 0$. Using binomial formula, $A^n = (1+x)^n = 1 + nx + y$, where y is a sum of positive numbers, so $y > 0$. Thus, $A^n > 1 + nx$ and so for each $\varepsilon > 0$ there exists such N that $A^n > 1/\varepsilon$ for all $n \geqslant N$ (this N can be found explicitly). Hence, $a^n < \varepsilon$ which means that $a^n \to 0$ when $n \to \infty$.

We can generalize the previous definition to sequences $(a_1, a_2, \ldots, a_n, \ldots)$ whose terms can also be negative. Then the numbers $|a_1|$, $|a_2|$, $\ldots$, $|a_n|$, $\ldots$ are nonnegative and we can apply the previous definition to them. We shall say that the sequence a_n approaches zero unboundedly, if the sequence of numbers $|a_n|$ unboundedly approaches 0. In that case one also writes $a_n \to 0$ when $n \to \infty$.

Now we have come to our main definition. If for a sequence $a = (a_1, a_2, \ldots, a_n, \ldots)$ there exists a number α, such that $a_n - \alpha \to 0$ when $n \to \infty$, then α is called the *limit* of the sequence a. One also says that the sequence a_n *tends* to α and one writes $a_n \to \alpha$ when $n \to \infty$.

Not every sequence has a limit. For example, if a sequence has a limit, then it is bounded. Really, let $\alpha_n \to \alpha$ when $n \to \infty$. Then there exists an N, such that $|\alpha_n - \alpha| < 1$ for $n > N$. Since $\alpha_n = \alpha + (\alpha_n - \alpha)$, it follows that $|\alpha_n| \leqslant |\alpha| + 1$ for $n > N$ and therefore $|\alpha_n| \leqslant C$ for all n, where C is the maximum of numbers $|\alpha_1|, \ldots , |\alpha_N|, |\alpha| + 1$. But even if a sequence is bounded, it can have no limit. An example is the sequence $(0, 1, 0, 1, \ldots)$ where 0 and 1 alternate. If it had a limit α, we could take in the definition of the limit $\varepsilon = \frac{1}{2}$ and we would have $|a_n - \alpha| < \frac{1}{2}$ for all $n > N$. But among a_n's with $n > N$ there are both 0 and 1. Therefore we would have $|\alpha| < \frac{1}{2}$ and $|1 - \alpha| < \frac{1}{2}$. Clearly, such a number α does not exist.

But if a sequence has a limit, this limit is unique. Namely, suppose that a sequence $(a_1, a_2, \ldots, a_n, \ldots)$ has two limits: α and β, $\alpha \neq \beta$. Then for each ε there exist numbers N and N', such that for $n > N$ it is $|a_n - \alpha| < \varepsilon$ and for $n > N'$ it is $|a_n - \beta| < \varepsilon$. Let $n > N$ and $n > N'$; then $|a_n - \alpha| < \varepsilon$ and $|a_n - \beta| < \varepsilon$, wherefrom $|\alpha - \beta| < 2\varepsilon$. But ε in our reasoning is an arbitrary positive number, and we can choose it so that $\varepsilon < \frac{1}{2}|\alpha - \beta|$, hence a contradiction.

As not every bounded sequence has a limit, considering just such sequences would not lead us to the construction of new real numbers. Our main result will be that there is a simple special type of sequences which always have limits and therefore they will give us a method of constructing new real numbers.

A sequence $(a_1, a_2, \dots, a_n, \dots)$ is called *increasing*, if $a_n \leqslant a_{n+1}$ for all n, i.e., $a_1 \leqslant a_2 \leqslant a_3 \leqslant a_4 \leqslant \cdots$.

Theorem 1. *Each bounded and increasing sequence of positive numbers has a limit.*

The proof will follow the logic of an anecdote which was popular when I was a student (i.e., before the war). The story was about different ways to catch a lion in a desert. There was a French method, method of NKVD-investigators, mathematician's method, ... Mathematician's method went like this. He divides the desert into two parts. The lion is situated in one of these parts. He divides this part again in two parts—and he continues like this till the lion appears in a part of the desert whose dimensions are less than the dimensions of the cage. It remains to put the cage around it. This was a parody to a way of proving existence theorems, one of which we are going to demonstrate now.

Let $a = (a_1, a_2, \dots, a_n, \dots)$ be an increasing sequence of positive numbers. By the assumption it is bounded, so there exists a constant C such that all $a_n < C$. Divide the segment $I_1 = [0, C]$ into two equal parts by the number $C/2$. Then one of the following is valid. Either there exists an m, such that $a_m \geqslant C/2$, and then all a_n with $n \geqslant m$ are contained in the segment $[C/2, C]$ (since the sequence is increasing); or $a_n \leqslant C/2$ for all n, and then all terms of the sequence belong to the segment $[0, C/2]$. Denote by I_2 one the segments, $[0, C/2]$ or $[C/2, C]$, namely the one which contains all the terms of sequence a, starting from some place. After that, divide the new segment into two parts. Obviously, we can continue the process unboundedly and we will obtain a sequence of embedded segments $I_1 \supset I_2 \supset I_3 \supset \cdots \supset I_m \supset \cdots$, where segment I_k has the length $C/2^k$, and which possesses the property that each segment I_k contains all the terms of sequence a, starting from some place. By the axiom of embedded segments (axiom VII) there exists a real number α, belonging to all the segments I_k. It is indeed the limit of sequence a. Really, as we have seen, all the terms of sequence a, starting from some place, belong to segment I_k, This means that for each natural number k there exists an N such that $a_n \in I_k$ for all $n > N$. But also $a \in I_k$. Since the length of segment I_k is equal to $C/2^k$, it follows that $|a_n - \alpha| < C/2^k$ for $n > N$. This gives us the property which appears in the definition of the limit, if we choose k so that $C/2^k < \varepsilon$. In particular, note that such a choice is always possible (the sequence $\left(1, \frac{C}{2}, \frac{C}{4}, \frac{C}{8}, \dots\right)$ tends to 0).

Theorem 1 is particularly useful when the sequence $a = (a_1, a_2, \dots, a_n, \dots)$ is the sequence of sums of a sequence of nonnegative numbers $c = (c_1, c_2, \dots, c_n, \dots)$ ($c_n \geqslant 0$), i.e., when $a_1 = c_1$, $a_2 = c_1 + c_2$, ... , $a_n = c_1 + c_2 + \cdots + c_n$. In such a case, obviously, the sequence a is increasing. But it has to be checked (and it could by no means be easy) whether it is bounded. For example, if in the sequence c all

$c_n = 1$, then $a_n = n$ and the sequence a is unbounded. We considered a less trivial example in Section 2 of Chapter IV: in the sequence c all $c_n = 1/n$. We saw that in that case the sequence a is also unbounded. But if we can check that the sequence a of sums is bounded, then according to Theorem 1 it has a unique limit α. This limit is called the *sum* of the sequence $(c_1, c_2, \dots, c_n, \dots)$, which is denoted by

$$c_1 + c_2 + \cdots + c_n + \cdots = \alpha.$$

Sometimes the infinite sum c is called a *series* and its sum—the *sum* of the series.

If the sequence of sums a_n is bounded, then, as we have seen, the sum of the series $c_1 + c_2 + \cdots + c_n + \cdots$ exists. If it is unbounded, then we say that the sum of the series does not exist. Hence, Lemma 1, Section 2, Chapter IV, states that the sum of the series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ does not exist.

Consider an example. Let a nonnegative number a, less than 1, be given, and let $c = (1, a, a^2, \dots, a^n, \dots)$. Then $a_n = 1 + a + a^2 + \cdots + a^{n-1}$ (in the n-th place in the sequence c there appears a^{n-1}). The sum $1 + a + a^2 + \cdots + a^{n-1}$ can be evaluated using the formula for the sum of a geometric progression—formula (12) of Chapter I:

$$a_n = 1 + a + a^2 + \cdots + a^{n-1} = \frac{1 - a^n}{1 - a} = \frac{1}{1 - a} - \frac{a^n}{1 - a}. \tag{1}$$

We have seen that $a^n \to 0$ when $n \to \infty$, wherefrom it follows immediately that $\frac{a^n}{1-a} \to 0$ when $n \to \infty$. Thus, formula (1) gives that $a_n \to \frac{1}{1-a}$. We can write this as:

$$1 + a + a^2 + \cdots + a^n + \cdots = \frac{1}{1 - a} \quad \text{for } a < 1. \tag{2}$$

The series on the left-hand side of relation (2) is called an *infinite geometric progression*, and formula (2) itself—the formula for the sum of an infinite geometric progression.

But there are examples of series where existence of sums is not hard to prove, but the explicit evaluation of the sums is much harder. For example, in Section 2 of Chapter IV we proved that the sums $\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2}$ are bounded. This means that the sum of the series $\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} + \cdots$ exists. But what is its value? This problem attracted mathematicians in the middle of XVIII century. It was Euler who solved it, when he found an interesting equality

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} + \cdots = \frac{\pi^2}{6}. \tag{3}$$

This was one of the most sensational Euler's discoveries. Euler went even further, evaluating the sum of the series $1 + \frac{1}{2^k} + \frac{1}{3^k} + \cdots + \frac{1}{n^k} + \cdots$ for arbitrary *even* k. It appeared that these sums were connected with the numbers of Bernoulli, which we described in the Appendix of Chapter II. Namely, the following formula is valid for each even k:

$$1 + \frac{1}{2^k} + \frac{1}{3^k} + \cdots + \frac{1}{n^k} + \cdots = \pi^k (-1)^{\frac{k}{2}-1} \frac{B_k}{2}\, k!. \tag{4}$$

We know nearly nothing about analogous sums with odd k. It was proved only recently (in 1978) that the sum $1+\frac{1}{2^3}+\frac{1}{3^3}+\dots+\frac{1}{n^3}+\cdots$ is an irrational number. This remains probably the only known fact about these sums for odd values of k.

Let us remark that just knowing the fact that a series $c_1+c_2+\dots+c_n+\cdots$ has a sum, one can deduce useful corollaries even if the value of the sum is not known.

Lemma 3. *If the sum of the series $c_1+c_2+\dots+c_n+\cdots$ exists, then the sequence of numbers $d_n=c_n+c_{n+1}+\cdots$ unboundedly approaches* 0.

We shall use an easy property of the limit. Suppose that a sequence a_1, a_2, $\dots$, a_n, $\dots$ has a limit α, i.e., $a_n\to\alpha$ when $n\to\infty$. Then for each number β the sequence $\beta-a_1$, $\beta-a_2$, $\dots$, $\beta-a_n$, $\dots$ has the limit $\beta-\alpha$. Really, the difference $\beta-\alpha-(\beta-a_n)=a_n-\alpha$, and the difference $a_n-\alpha\to 0$, hence $\beta-\alpha-(\beta-a_n)\to 0$ when $n\to\infty$. Denote the sum of the series $c_1+c_2+\dots+c_n+\cdots$ by α and the number $c_1+c_2+\dots+c_m$ by a_m. By the definition of the sum of an infinite series, the sum α of the series $c_1+c_2+\dots+c_n+\cdots$ is equal to the limit of the sequence $a_1, a_2, \dots, a_m, \dots$. In the same way the sum d_n of the sequence $c_{n+1}+c_{n+2}+\cdots$ is equal to the limit of the sequence $a_{n+1}-a_n$, $a_{n+2}-a_n$, $\dots$ $a_{n+k}-a_n$, $\dots$. By the remark from the beginning of the proof, the last limit is equal to $\alpha'-a_n$, where α' is the limit of the sequence a_{n+1}, a_{n+2}, $\dots$, a_{n+k}, $\dots$ (for fixed n). But the limit of the sequence a_{n+1}, a_{n+2}, $\dots$ is the same as the limit of the sequence a_1, a_2, $\dots$, i.e., $\alpha'=\alpha$. We obtain that $d_n=\alpha-a_n$. But, by the definition of limit, $\alpha-a_n\to 0$, i.e., $d_n\to 0$ when $n\to\infty$.

As an example, put $d_n=\frac{1}{n^2}+\frac{1}{(n+1)^2}+\cdots$. We see that $d_n\to 0$ when $n\to\infty$.

Considering limits of infinite sums leads us away from algebra, which is mainly concerned with finite expressions. These questions are closely related with another branch of mathematics, called analysis. That is why we are not going to consider them in more detail. Let us remark only that the most interesting results—such as formulas (3) and (4)—appear on borders of these areas.

Problems

1. Prove that if the sum of the series $c_1+c_2+\dots+c_n+\cdots$ exists, then $c_n\to 0$ when $n\to\infty$.

2. Prove that if $a_n<C$ for each n and $a_n\to\alpha$ when $n\to\infty$, then $\alpha\leqslant C$. Give an example when equality is obtained.

3. Let $a_n\to\alpha$ when $n\to\infty$. Put $b_n=a_{2n}$. Does the sequence b_1, b_2, $\dots$ have a limit and what is its value? Is it possible, from the existence of the limit of this sequence, to conclude that the sequence a_1, a_2, $\dots$ itself has a limit? If it does have a limit, what is its value?

4. Does there exist a limit of the sequence a_1, a_2, ... where

$$a_n = \frac{1}{2} - \frac{1}{3} + \cdots + \frac{(-1)^n}{n}\,?$$

Hint. Group consecutive terms in pairs.

5. Let $f(x)$ be a polynomial of degree d. Prove that $a_n \to 0$ when $n \to \infty$, where $a_n = f(n)/n^{d+1}$.

6. Find the sum of the series $b + ba + ba^2 + ba^n + \cdots$, where $|a| < 1$ and b is arbitrary. Usually, the sequence $(b, ba, ba^2, \dots)$ is also called an infinite geometric progression.

7. In a square with side b, centres of the sides are joined by segments. In the new square which is obtained in that way the same procedure is done, etc. Find the sum of areas of all squares that can be obtained in this way.

8. Find the sum of the series $\dfrac{1}{1\cdot 2} + \dfrac{1}{2\cdot 3} + \cdots + \dfrac{1}{n\cdot(n+1)} + \cdots$.

9. Construct a sequence of positive rational numbers smaller than 1, such that a_n has the denominator n and which does not have a limit.

10. Prove that if the sequence a_1, a_2, ... has a limit α, and the sequence b_1, b_2, ... has a limit β, then the sequence of sums $a_1 + b_1$, $a_2 + b_2$, ... has the limit $\alpha + \beta$.

3. Decimal representation of real numbers

In Section 1 we described real numbers using a system of axioms. Now we are going to show how real numbers can be given concretely. Here we shall not say anything new—we shall speak about justification of the well known representation of real numbers by infinite decimal fractions. But now we shall show how the existence of such a representation can be deduced from axioms listed in Section 1.

We shall use the usual representation in which integer part can be either positive or negative, while fractional part (sometimes called the mantissa) is always nonnegative.

Let A be an arbitrary integer (of either sign) and $a_1, a_2, \dots, a_n, \dots$ an infinite sequence of numbers, each of which can take one of 10 values: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. All this together will be denoted by $A, a_1a_2a_3\dots$ and called an infinite decimal fraction. For the time being it is just an infinite sequence, written in a different way. Now we are going to show how a real number can be corresponded to it. We define, for each index n, a number

$$\alpha_n = A + \frac{a_1}{10} + \cdots + \frac{a_n}{10^n}. \tag{5}$$

Obviously, the sequence $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ is increasing. Let us prove that it is bounded. Really, since all $a_i \leqslant 9$, we have

$$\frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_n}{10^n} \leqslant \frac{9}{10}\left(1 + \frac{1}{10} + \cdots + \frac{1}{10^{n-1}}\right).$$

We apply the formula about the sum of geometric progression:

$$1+\frac{1}{10}+\cdots+\frac{1}{10^{n-1}}=\frac{1-\frac{1}{10^n}}{1-\frac{1}{10}}<\frac{10}{9}$$

and as a result we obtain that

$$\frac{a_1}{10}+\frac{a_2}{10^2}+\cdots+\frac{a_n}{10^n}<1 \tag{6}$$

so that $\alpha_n < A+1$.

By Theorem 1, the sequence $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ has a limit α. Real number α will be called the number *corresponded* to the infinite decimal fraction, and this will be denoted by

$$\alpha = A, a_1 a_2 \dots a_n \dots \tag{7}$$

Sometimes it is said that α is *equal* to the decimal fraction $A, a_1 a_2 \dots a_n \dots$. This simply means that α is equal to the sum of the infinite series $A+\frac{a_1}{10}+\cdots+\frac{a_n}{10^n}+\cdots$.

Our next goal is to explore this correspondence between decimal fractions and real numbers. Is it bijective? In other words: can a real number correspond to two different decimal fractions? And is each real number corresponded to some decimal fraction?

Consider the first question. First of all, remark that the answer is sometimes positive. Take, e.g., the infinite decimal fraction $0,9999\dots$, where each decimal after the comma is equal to 9. Which real number does it represent? According to general definition we have to consider the sequence $\alpha_n=\frac{9}{10}+\frac{9}{10^2}+\cdots+\frac{9}{10^n}$. This sum is easy to evaluate: according to the formula about the sum of geometric progression (formula (12) in Chapter I) it is equal to

$$\frac{9}{10}\left(1+\frac{1}{10}+\cdots+\frac{1}{10^{n-1}}\right)=\frac{9}{10}\frac{1-\frac{1}{10^n}}{1-\frac{1}{10}}=\frac{9}{10}\frac{1-\frac{1}{10^n}}{\frac{9}{10}}=1-\frac{1}{10^n}.$$

Obviously, the limit of the sequence $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ is equal to 1, so that $1=0,9999\dots$. But, on the other hand, surely $1=1,00\dots$, where in front of the comma there is just 1, and after it all zeros. In such a way, the same real number 1 is corresponded to two distinct infinite decimal fractions.

It is clear that one can construct a lot of examples of the same kind. In general, such an example has the following form. Let an infinite decimal fraction has the form $A, a_1 \dots a_k 99\dots$, i.e., suppose that starting form some place (in our case from the $(k+1)$-st one) all the decimals are equal to 9. We can assume that $a_k \neq 9$, i.e., k-th is the first place after which all the 9's follow. Then, literally repeating previous reasoning, one can conclude that this fraction is equal to the same number as the fraction $A, a_1 \dots a_{k-1}(a_k+1)000\dots$, in which all the decimals after the k-th one are equal to 0. A fraction having all the decimals 9, starting from some place, is said to have 9 as a period. We have seen that for such fractions one-to-one correspondence between fractions and real numbers is violated.

It is a bit of a surprise that such violation appears only in those cases.

Theorem 2. *Two distinct infinite decimal fractions, neither of which has 9 as a period, are corresponded to distinct real numbers.*

The proof can be obtained easily if we connect our construction of a real number, defined by a decimal fraction, with the usual measuring of numbers with accuracy of $1/10^m$, with deficiency and excess. One has to divide the line into segments of the length $1/10^m$, whose endpoints are rational numbers with denominator 10^m. Then each point from the line, that is, each real number, falls in one of the segments. The endpoints of the segment give a measure of the number, with deficiency and excess and accuracy of $1/10^m$. However, violating of one-to-one correspondence appears because of the endpoints of segments themselves. To which of the segments, left or right, is each of these points corresponded? This is the same problem which appears in connection with number 9 in the period. We are going to show that our choice (without 9 in periods) corresponds to the case when the endpoints of segments are always attached to segments on the right-hand side. In other words, the constructed numbers α_m and the number α which they define are connected by the relation

$$\alpha_m \leqslant \alpha < \alpha_m + \frac{1}{10^m}. \tag{8}$$

(The fact that numbers α_m are rational with the denominators of the form 10^m follows from their form (5).)

Remember that number α was defined as the limit of the sequence $\alpha_1, \alpha_2, \ldots,$ $\alpha_n, \ldots$. All numbers α_n with $n \geqslant m$, obviously satisfy the condition $\alpha_n \geqslant \alpha_m$. Hence, such an inequality is valid for their limit α. Really, from the assumption $\alpha < \alpha_m$ we could deduce that $\alpha_n - \alpha = (\alpha_n - \alpha_m) + (\alpha_m - \alpha) \geqslant \alpha_m - \alpha$ for all $n \geqslant m$. But, by the definition of limit, the absolute value of the number $\alpha_n - \alpha$ is smaller than an arbitrary given positive number for n large enough. This contradicts the fact that it is not smaller than the fixed positive number $\alpha_m - \alpha$ (see Problem 2 in Section 2).

In this way the left-hand inequality in (8) is proved. The right-hand one can be proved similarly, if the sign $<$ is replaced by $\leqslant$. Namely, for each $n > m$ we have

$$\alpha_n = \alpha_m + \frac{a_{m+1}}{10^{m+1}} + \cdots + \frac{a_n}{10^n} \leqslant \alpha_m + \frac{1}{10^m}\left(\frac{a_{m+1}}{10} + \cdots + \frac{a_n}{10^{n-m}}\right) \tag{9}$$

and applying inequality (6) we conclude that $\alpha_n < \alpha_m + \frac{1}{10^m}$. Repeating the previous reasoning we obtain that $\alpha \leqslant \alpha_m + \frac{1}{10^m}$.

But, if we want to obtain the right-hand inequality in (8) with the sign $<$, we have to use the fact that the fraction $A, a_1 a_2 \ldots$ does not have 9 as a period. The proof is only a bit more complicated. Let us prove the right-hand inequality in (8) for fixed index m. We shall use the fact that the decimal fraction does not have 9 as a period. That means that somewhere after a_m there has to appear a digit a_k different from 9. For an arbitrary $n > k$ we can write

$$\alpha_n = \alpha_m + (a_{m+1}/10^{m+1} + \cdots + a_k/10^k) + (a_{k+1}/10^{k+1} + \cdots + a_n/10^n).$$

As before, we see that

$$a_{k+1}/10^{k+1} + \cdots + a_n/10^n \leqslant 1/10^k$$

and so

$$\alpha_n \leqslant \alpha_m + (a_{m+1}/10^{m+1} + \cdots + (a_k + 1)/10^k).$$

Since $a_k \neq 9$, the digit $a_k + 1$ is one of the digits 1, 2, ... , 9. Put

$$c = a_{m+1}/10 + \cdots + (a_k + 1)/10^{k-m}.$$

We can repeat our reasoning once more and obtain that $c < 1$. Number c depends only on the choice of m and k, and not on n. Hence, replacing α_n by its limit α, we obtain, as before, $\alpha \leqslant \alpha_m + c/10^m < \alpha_m + 1/10^m$.

That proves inequality (8).

It follows right away from the inequality (8) that to each two distinct decimal fractions, not having 9 as a period, there correspond two distinct real numbers. Let, to the contrary, the same number α corresponds to fractions $A, a_1a_2\ldots$ and $A', a_1'a_2'\ldots$. Then together with inequalities (8) we have relations

$$\alpha_m' \leqslant \alpha < \alpha_m' + \frac{1}{10^m},$$

where $\alpha_m' = A' + \dfrac{a_1'}{10} + \cdots + \dfrac{a_m'}{10^m}$. Let $\alpha_m' \neq \alpha_m$ and $\alpha_m' > \alpha_m$. From these relations it follows that $\alpha_m' < \alpha_m + \frac{1}{10^m}$, i.e., $\alpha_m' - \alpha_m < \frac{1}{10^m}$. But this contradicts the fact that α_m and α_m' are distinct rational numbers having the same denominator 10^m. Hence, $\alpha_m' = \alpha_m$ for all m. But numbers a_m are uniquely determined by the numbers α_m, since $\alpha_m - \alpha_{m-1} = a_m/10^m$. Thus, they coincide in both fractions, too.

We pass now to the second question: does every real number correspond to some infinite decimal fraction? As well as the answer, the method of proof is already known to us. We just want to convince ourselves that the reasoning can be based on the axioms we formulated.

First of all, let us remark that each real number α is situated between two consecutive integers, i.e., there exists an integer A, such that $A \leqslant \alpha < A + 1$. Let, for start, α be positive. Applying Archimedes' axiom, we conclude that there is an integer n such that $\alpha < n$. Obviously, $n > 0$, and since there exist only a finite number of natural numbers not exceeding n, there also exists the last (the smallest) one with that property. Denote this number by m. Then $\alpha < m$, but $m - 1$ does not possess this property; that means $m - 1 \leqslant \alpha < m$ and $A = m - 1$ has the desired properties. If α is negative, we put $\alpha' = -\alpha$. Then $\alpha' > 0$ and we can apply our procedure: there exists n such that $n \leqslant \alpha' < n + 1$. Axiom IV$_3$ implies that $-(n + 1) < \alpha \leqslant -n$. If $\alpha' \neq n$, we can put $A = -(n + 1)$ and $A < \alpha < A + 1$. If $\alpha' = -n$, then we have to put $A = -n$. And so, for each real number α there exists an integer A such that $A \leqslant \alpha < A + 1$, hence α can be represented as $\alpha = A + \varepsilon$, where $0 \leqslant \varepsilon < 1$.

Now observe that if some three numbers a_1, a_2, a_3 satisfy $a_1 < a_2$ and $a_2 < a_3$, then for each α satisfying conditions $a_1 \leqslant \alpha < a_3$, one of the following conditions must be satisfied: either $a_1 \leqslant \alpha < a_2$ or $a_2 \leqslant \alpha < a_3$. The fact is demonstrated in Fig. 1 where the interval $[a_1, a_3)$ is simply the union of the intervals $[a_1, a_2)$ and $[a_2, a_3)$. Formally, it is a consequence of the fact that for each α exactly one of the relations $\alpha < a_2$, $a_2 < \alpha$ and $a_2 = \alpha$ holds.

Fig. 1

Consider a more general case. Let the following conditions be satisfied for n numbers $\alpha_1, \dots, \alpha_n$: $\alpha_1 < \alpha_2$, $\alpha_2 < \alpha_3$, $\dots$, $\alpha_{n-1} < \alpha_n$. Then for each number α, satisfying $\alpha_1 \leqslant \alpha < \alpha_n$, one of the conditions $\alpha_{i-1} \leqslant \alpha < \alpha_i$ $(i = 2, 3, \dots, n)$ is valid. In order to prove it one just has to apply the previous assertion to the case of three numbers α_1, α_2, α_n. Then either $\alpha_1 \leqslant \alpha < \alpha_2$ (and our statement is valid for $i = 2$), or $\alpha_2 \leqslant \alpha < \alpha_n$. In the latter case consider numbers α_2, α_3, α_n, etc. For some i we come to the desired condition $\alpha_{i-1} \leqslant \alpha < \alpha_i$.

We can return now to our original question. We have already proved that each real number α can be represented in the form $A + \varepsilon$, where A is an integer and $0 \leqslant \varepsilon < 1$. Consider now numbers $\dfrac{k}{10}$, $k = 0, 1, \dots, 10$. According to the previous result, we can conclude that $\dfrac{k}{10} \leqslant \varepsilon < \dfrac{k+1}{10}$ for some k, $0 \leqslant k < 10$. Denoting this number by a_1, we can write $\varepsilon = \dfrac{a_1}{10} + \varepsilon_1$, where $0 \leqslant \varepsilon_1 < \dfrac{1}{10}$. Hence, $\alpha = A + \dfrac{a_1}{10} + \varepsilon_1$. Continuing the process, we obtain numbers $a_1, \dots, a_n, \dots$, where always $0 \leqslant a_i \leqslant 9$, and the sequence $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$, where $\alpha_n = A + \dfrac{a_1}{10} + \dots + \dfrac{a_n}{10^n}$, has the limit α, i.e., the number α is corresponded to the infinite decimal fraction $A, a_1 a_2 \dots a_n \dots$.

Summing up, one can say that *forming infinite decimal fractions for real numbers does not establish a one-to-one correspondence between infinite decimal fractions and real numbers, but such a correspondence becomes one-to-one if we exclude those decimal fractions which have 9 as a period.*

Problems

1. Prove that a real number α corresponds to an infinite decimal fraction having 0 as a period if and only if α is a rational number a/b where a and b are integers such that just 2 and 5 can be prime factors of b.

2. When finding the infinite decimal fraction which corresponds to a rational number a/b, it is enough to find the mantissa, so we can assume that $0 < a < b$. Let $\alpha_n = \dfrac{a_1}{10} + \dfrac{a_2}{10^2} + \dots + \dfrac{a_n}{10^n}$, where $0, a_1 a_2 \dots$ is the infinite decimal fraction corresponding to the number a/b. Prove that $\dfrac{a}{b} - \alpha_n = \dfrac{r_n}{10^n b}$, where $0 \leqslant r_n < b$ and the numbers r_n are connected by the relation $10r_{n-1} = ba_n + r_n$, i.e., a_n is the quotient and r_n the remainder when $10r_{n-1}$ is divided by b. Convince yourself that this method of successive evaluation of digits a_n of a decimal fraction agrees with the usual division algorithm.

3. Prove that the infinite decimal fraction corresponding to a rational number is periodic, i.e., it has the form $(** \cdots)(\mathcal{P})(\mathcal{P})\cdots$, where $(** \cdots)$ denotes a certain finite group of symbols, after which the group of symbols $(\mathcal{P})$, called the *period*, repeats. *Hint.* Use Problem 2 (i.e., the division algorithm) and note that the possible number of remainders when $10r_{n-1}$ is divided by b is finite (not greater than b).

4. Prove that if the denominator b of the fraction a/b is relatively prime with 10, then the period begins immediately after the comma.

5. Under the assumptions of Problem 4, prove that the number of digits in the period is equal to the smallest number k for which $10^k - 1$ is divisible by b.

6. Under the assumptions of Problems 4 and 5, prove that the number of digits in the period is not greater than the number of natural numbers not exceeding b and relatively prime with b. This number is given by formula (25) of Chapter III.

7. Prove that each periodic infinite decimal fraction corresponds to a rational number A. Namely, if $A, a_1 a_2 \ldots a_n$ stays in front of the period $(p_0, p_1, \ldots, p_{m-1})$, and $A + \dfrac{a_1}{10} + \cdots + \dfrac{a_n}{10^n} = Q$, $p_0 10^{m-1} + p_1 10^{m-2} + \cdots + p_{m-1} = \mathcal{P}$, then the rational number corresponding to the given fraction is $Q + \dfrac{\mathcal{P}}{10^n(10^m - 1)}$.

8. Prove that the infinite decimal fraction $0{,}1010010001\ldots$, where the number of zeros between two consecutive 1's increases by 1 each time, corresponds to an irrational number.

4. Real roots of polynomials

Having made a firmer basis for the theory of real numbers, we can now obtain some new results about real roots of polynomials with real coefficients. In order to do this, we have to investigate first the behaviour of a polynomial $f(x)$ in the neighbourhood of a value $x = a$.

Theorem 3. *For each polynomial $f(x)$ and each number a there exists a constant M, such that the inequality*

$$|f(x) - f(a)| \leqslant M|x - a| \tag{10}$$

is valid for all x such that $|x - a| \leqslant 1$.

Remember that $|A|$ (read as "absolute value of number A"), by the definition, is equal to A if $A \geqslant 0$ and to $-A$ if $A < 0$. It follows that $|A|$ is always a nonnegative number. From school courses it is known that

$$|A+B| \leqslant |A|+|B| \tag{11}$$

$$|A+B| \geqslant |A|-|B| \tag{12}$$

$$|AB| = |A|\cdot|B|. \tag{13}$$

Theorem 3 gives a quantitative estimate of how much $f(x)$ differs from $f(a)$ if x slightly differs from a. In order to prove the theorem, put $y=x-a$, i.e., $x=a+y$ and substitute this value into the polynomial $f(x)$. Each term $a_k x^k$ of the polynomial $f(x)$, after the substitution, gives the expression $a_k(a+y)^k$, which can be written as a sum of powers of y and then similar terms in $f(a+y)$ can be reduced. As a result we obtain that $f(a+y)$ is a polynomial in y, which we denote by $g(y)=c_0+c_1y+\cdots+c_ny^n$. Then $f(x)=f(a+y)=g(y)$, $f(a)=f(a+0)=g(0)=c_0$, $x-a=y$ and inequality (10) which we intend to prove becomes

$$|g(y)-g(0)| \leqslant M|y| \tag{14}$$

for all y satisfying $|y|\leqslant 1$.

In the transformed form, the expression $g(y)-g(0)$ acquires a simple form $c_1y+\cdots+c_ny^n$ (since $g(0)=c_0$). Inequality (11) can be applied also to a sum with an arbitrary number of summands (which can be proved directly by induction) and, in particular, to our sum $c_1y+\cdots+c_ny^n$. We obtain that

$$|g(y)-g(0)| = |c_1y+\cdots+c_ny^n| \leqslant |c_1y|+\cdots+|c_ny^n|.$$

Using equality (13) (also applied to an arbitrary number of factors), $|c_ky^k|=|c_k|\cdot|y|^k$, so that

$$|g(y)-g(0)| \leqslant |c_1||y|+\cdots+|c_n||y|^n.$$

Since, by the assumption, $|y|\leqslant 1$, we have $|y|^k\leqslant|y|$ and

$$|g(y)-g(0)| \leqslant (|c_1|+\cdots+|c_n|)|y|$$

for $|y|\leqslant 1$.

It is enough to put $M=|c_1|+\cdots+|c_n|$ to obtain inequality (14), which also means inequality (10).

Now we are able to prove an important property of polynomials.

Theorem 4. (Bolzano's theorem) *If a polynomial for $x=a$ and $x=b$ takes values with opposite signs, then it takes the value 0 somewhere between a and b.*

In other words, if for a polynomial $f(x)$ values $f(a)$ and f(b) are numbers of opposite signs and $a<b$, then there exists c, such that $a<c<b$ and $f(c)=0$.

Theorem 4 appears rather obvious if one looks at the graph of the polynomial $f(x)$ (Fig. 2). It states that the graph cannot "jump" across the x-axis without intersecting it. On the other hand, it is completely possible to *draw* such a graph (Fig. 3). So, we have to prove that such a graph cannot be the graph of a polynomial. For more general functions it is connected with a rather involved property

Fig. 2

Fig. 3

which is called *continuity*. In the case of polynomials it is enough to use the easy inequality (10), proved in Theorem 3.

The proof is based on the same principle of "catching a lion in a desert", we have already used for proving Theorem 1.

Suppose, for example, that $f(a) > 0$, $f(b) < 0$. Consider the segment $[a, b]$ (i.e., the set of real numbers x satisfying $a \leqslant x$ and $x \leqslant b$). Denote this segment by I_1 and divide it into two segments of equal length by the point $r = \frac{a+b}{2}$. If $f(r) = 0$, then the theorem is proved ($c = r$). If $f(r) \neq 0$ and, for example, $f(r) > 0$, then the polynomial $f(x)$ takes values of opposite signs for $x = r$ and $x = b$. Denote then by I_2 the segment $[r, b]$. If that $f(r) < 0$, then the segment $[a, r]$ will be denoted by I_2. In any case we obtain a segment I_2 contained in I_1, having two times smaller length, and having again the property that the polynomial $f(x)$ has values of opposite signs at its endpoints—namely, positive at the left-hand end and negative at the right-hand one.

This process can be continued. Either we shall at some moment reach a root of the polynomial $f(x)$ (and the theorem will be proved), or the process shall continue unboundedly. It remains to consider the latter case. We obtain an infinite sequence of embedded segments $I_1 \supset I_2 \supset \cdots \supset I_n \supset \cdots$, $I_n = [a_n, b_n]$, such that each of them is of half-a-length of the previous one, and the polynomial $f(x)$ takes values of opposite signs at the endpoints a_n and b_n of each segment I_n, more precisely, $f(a_n) > 0$, $f(b_n) < 0$. Now we are going to use the more precise definition of real numbers we gave in Section 1. Segments I_n satisfy the prepositions of Axiom VII (axiom of embedded segments) and Lemma 1 of Section 1. Really, segments I_n are embedded one into another, by their construction, and since I_n is half-of-length of segment I_{n-1}, its length is equal to $\dfrac{b-a}{2^{n-1}}$, and so this length becomes unboundedly small when n increases. Hence, according to Axiom VII and Lemma 1, there exists a unique number c, belonging to all segments I_n, i.e., such that

$$a_n \leqslant c \leqslant b_n. \tag{15}$$

In this way we have constructed the number c which we searched for. Namely, we now prove that $f(c) = 0$.

Consider the values $f(a_n)$ of the polynomial $f(x)$ at the left-hand endpoints of segments I_n. By the assumption, all $f(a_n) > 0$. Inequality (11) implies that the sequence $a_1, a_2, \ldots$ approaches the number c unboundedly: really, $a_n \leqslant c \leqslant b$ and

$0 \leqslant c - a_n \leqslant b_n - a_n$, where, by the assumption, $b_n - a_n = \dfrac{b-a}{2^{n-1}}$. Therefore the inequality $|a_n - c| < \varepsilon$ will be satisfied if $\dfrac{b-a}{2^{n-1}} < \varepsilon$, and this will be valid for each $\varepsilon > 0$ if n is chosen large enough. Let us prove that it follows from this that the values $f(a_n)$ approach the value $f(c)$ unboundedly. Really, in order to prove that $|f(a_m) - f(c)| \leqslant \varepsilon$ for m large enough, we can use inequality (10) from Theorem 3. Since a_m approaches c unboundedly, we have $|a_m - c| < 1$ for m large enough, and we can apply inequality (10). We see that $|f(a_m) - f(c)| < M|a_m - c|$ and so $|f(a_m) - f(c)| < \varepsilon$ if $M|a_m - c| < \varepsilon$, i.e., if $|a_m - c| < \varepsilon/M$. But we have convinced ourselves that this inequality is valid for m large enough (since ε/M can again be denoted by ε!).

What can be said about the number $f(c)$, which is known to be the limit of the sequence of positive numbers $f(a_n)$? Clearly, $f(c) \geqslant 0$. Really, if $f(c)$ were negative, than for positive $f(a_n)$ we would have $f(a_n) - f(c) \geqslant -f(c)$, and hence $|f(a_n) - f(c)| \geqslant -f(c)$, but this would contradict the fact that $|f(a_n) - f(c)| < \varepsilon$ if $\varepsilon < -f(c)$.

We have thus proved that $f(c) \geqslant 0$. Following exactly the same arguments, considering numbers b_n satisfying $f(b_n) < 0$, we can prove that $f(c) \leqslant 0$. Therefore, for the number $f(c)$ only one possibility remains—$f(c) = 0$. The theorem is proved.

One should pay attention to a completely new way of reasoning in proving this theorem. We have proved in fact (under certain conditions) the existence of a root of the polynomial $f(x)$. But we have not done it using any kind of formula (as, for example, when solving a quadratic equation) but using the axiom of embedded segments. But, at the same time, it is by no means a pure “theorem of existence”, where we know only that a certain quantity exists—and nothing more than that. For example, we can in fact find the root c with deficiency and excess and with an arbitrary prescribed accuracy, constructing numbers a_n and b_n such that c lies between them (inequality (15)) and which get closer and closer to each other.

Bolzano’s theorem gives us the possibility to know a lot about concrete polynomials. Consider, for example, the polynomial $f(x) = x^3 - 7x + 5$ and make a table of its values for integer values of x, with small absolute values (Table 1). One can see from the table that the polynomial $f(x)$ takes values of opposite signs at the ends of the segments $[2, 3]$, $[0, 1]$ and $[-3, -2]$. By Bolzano’s theorem it has a root in each of these segments. Hence, the polynomial $f(x)$ has at least three roots. But its degree is equal to 3 and by Theorem 3 of Chapter II it cannot have more than 3 roots. We have proved that the polynomial $f(x)$ has exactly 3 roots and they lie in segments $[2, 3]$, $[0, 1]$ and $[-3, -2]$.

x	-3	-2	-1	0	1	2	3
$f(x)$	-1	11	11	5	-1	-1	11

Table 1.

There are some other polynomials for which Bolzano’s theorem gives the precise answer, too. An important case is the polynomial $x^n - a$ whose roots are called “roots of a of degree n” (denoted as $\sqrt[n]{a}$). Consider first the case when $a > 0$. Then the polynomial $f(x) = x^n - a$ takes for $x = 0$ negative value $-a$. On the other hand, it is easy to find a value $x = c$ such that $f(c) > 0$ Really, by Archimedes' axiom (Axiom VI) there exists a natural number m such that $m > a$. Then $m^n > m$ and $m^n - a > m - a > 0$. Using Bolzano's theorem, we can state that there is a root of the polynomial in the segment $[0, m]$. If, on the other hand, $a < 0$ and n is even, then such polynomials obviously do not have roots: $x^n \geqslant 0$ as an even power of a real number, and $x^n - a > 0$. If n is odd, then putting $x = -y$ we obtain that $x^n - a = -y^n - a = -(y^n + a)$. The polynomial $y^n + a$ (for $a < 0$), as we have just proved, has a root, and so the same is true for the polynomial $x^n - a$. In school courses these arguments are usually omitted (because of the lack of a precise theory of real numbers), but it is proved (very easily) that for n odd the polynomial $x^n - a$ does not have more than one root (as we have seen—it has exactly one) and that for n even and $a > 0$—not more than two roots which differ only in the sign (which means it has exactly two roots).

But in the case of other polynomials, it can happen that Bolzano's theorem does not give anything. Take as an example the polynomial $x^2 - x + 2$. Using the formula for solutions of quadratic equation we can conclude that this polynomial has no real roots. But if we tried to give values $0, \pm 1, \pm 2, \ldots$ to the argument x, we would obtain only positive values, and Bolzano's theorem wouldn't give us anything. Therefore, we will try now to explore polynomials more thoroughly.

Theorem 3 estimates values of a polynomial for values of x being close to a certain value a. We shall prove now a similar assertion about values of the polynomial for large (by absolute value) values of x.

Theorem 5. *For the polynomial $f(x) = a_0 + a_1x + \cdots + a_nx^n$ there exists a constant $N > 0$ such that*

$$|a_0 + a_1x + \cdots + a_{n-1}x^{n-1}| < |a_nx^n| \tag{16}$$

for all values of x such that $|x| > N$.

The theorem states that for sufficiently large values of x, the absolute value of the leading term exceeds the absolute value of the sum of all other terms. In order to prove this, we use inequality (11) (for an arbitrary number of summands) and equality (13). It follows from them that $|a_0 + a_1x + \cdots + a_{n-1}x^{n-1}| \leqslant |a_0| + |a_1||x| + \cdots + |a_{n-1}||x|^{n-1}$, and $|a_nx^n| = |a_n||x|^n$. In order to prove inequality (16) it is enough to convince oneself that $|a_0| + |a_1||x| + \cdots + |a_{n-1}||x|^{n-1} \leqslant |a_n||x|^n$, and this will be proved if we show that

$$|a_k||x|^k < \frac{1}{n}|a_n||x|^n \tag{17}$$

for each $k = 0, 1, \ldots, n-1$ and $|x| > N$ for N large enough. Then, summing up all the inequalities (17) for $k = 0, 1, \ldots, n-1$ we obtain the inequality we needed.

Inequality (17) can be solved in the usual way. It is equivalent to

$|x|^{n-k} > \dfrac{n|a_k|}{|a_n|}$, i.e.,

$$|x| > \sqrt[n-k]{n\frac{|a_k|}{|a_n|}}. \tag{18}$$

Therefore, it is enough to choose for N an arbitrary number larger than all the numbers $\sqrt[n-k]{n\dfrac{|a_k|}{|a_n|}}$, $k = 0, 1, \dots, n-1$, and it will satisfy the assertion of Theorem 5.

Theorem 5 has a lot of useful corollaries. Note first that under the assumptions of the theorem (i.e., for $|x| > N$) we always have $|f(x)| > 0$, which follows immediately from inequality (12)

$$|f(x)| = |a_0 + a_1x + \dots + a_nx^n| \leqslant |a_nx^n| - |a_1 + a_1x + \dots + a_{n-1}x^{n-1}|.$$

But this means that the polynomial $f(x)$ does not have roots x with $|x| > N$. In other words, roots of a polynomial (if they exist) have to be contained in the segment $|x| \leqslant N$, where, as we have shown (inequality (18)) N can be chosen as the greatest of the numbers $\sqrt[n-k]{n\dfrac{|a_k|}{|a_n|}}$. One calls such a number N the *bound of roots* of the polynomial. So, for the polynomial $x^3 - 7x + 5$ one can take for N an arbitrary number greater than $\sqrt[3]{3 \cdot 5}$ and $\sqrt{3 \cdot 7}$. For example, $N = 4{,}6$ satisfies the conditions. This means that all roots of the polynomial are distributed between $-4{,}6$ and $4{,}6$. We have convinced ourselves earlier that they are in fact contained between -3 and $+3$ (Table 1).

Theorem 5 implies more that just the assertion that $f(x) \neq 0$ if $|x| > N$, for the found value of N. To evaluate the value of $a_0 + a_1x + \dots + a_{n-1}x^{n-1} + a_nx^n$ means to sum up two real numbers $a_0 + a_1x + \dots + a_{n-1}x^{n-1}$ and a_nx^n, first of which is smaller (by absolute value) than the other (for $|x| > N$). But then the sign is determined by the sign of the second summand. We come to the following conclusion:

Corollary 1. *For $|x| > N$, where N is the bound of roots defined in Theorem 5, values of the polynomial $f(x)$ have the same sign as the leading term a_nx^n.*

Suppose that the degree n of the polynomial is odd. Then the sign of the leading term a_nx^n for $x > 0$ agrees with the sign of the coefficient a_n, and for $x < 0$ it is opposite. Corollary 1 shows that for $x > N$ and $x < -N$ the polynomial itself acquires values of opposite signs (namely, the signs of a_n and of $-a_n$). Bolzano's theorem implies that between these values there is at least one root of the polynomial. We obtained the following proposition:

Corollary 2. *Each polynomial of odd degree has at least one root.*

This is really an unexpected result. In fact, you know that a polynomial of the second degree may have no roots (e.g., the polynomial $x^2 + 1$). One may think that the same could happen to polynomials of greater degrees: 3, etc. But here, according to the corollary, a polynomial of the third degree always has a root. The situation appears more complicated; it depends not on how large the degree of the polynomial is, but on its parity.

Finally, consider one more property of polynomials, which can make investigations in some cases much easier. Theorem 3 gave us information about the absolute value of the difference $f(x) - f(a)$ when the difference $x - a$ is small. We shall investigate now the *sign* of the difference $f(x) - f(a)$. Here we shall exclude the cases when the value $x = a$ appears to be a root of the derivative $f'(x)$ of the polynomial $f(x)$. These special values of a could be investigated easily in the same manner, but we will not need this at the moment.

Theorem 6. *Let a polynomial $f(x)$ be given and take a value $x = a$ which is not a root of its derivative $f'(x)$ (i.e., $f'(a) \neq 0$). If $f'(a) > 0$, then the values $f(x)$ for x close, but to the left of a, are smaller than $f(a)$, and for x close, but right of a, are greater than $f(a)$. If $f'(a) < 0$, then the situation is opposite.*

$f'(a) > 0$

Fig. 4

$f'(a) < 0$

Fig. 5

This means that there exists sufficiently small $\varepsilon > 0$ (depending on $f(x)$ and on a), such that when $f'(a) > 0$, for $a - \varepsilon < x < a$ we have $f(x) < f(a)$, and for $a < x < a + \varepsilon$, we have $f(x) > f(a)$. If, however, $f'(a) < 0$, then for $a - \varepsilon < x < a$ we have $f(x) > f(a)$, and for $a < x < a + \varepsilon$, we have $f(x) < f(a)$ (see graphs of $f(x)$ on Figs. 4 and 5).

The proof is quite easy. We know by Bezout's theorem that the polynomial $f(x) - f(a)$ is divisible by $x - a$. Therefore

$$f(x) - f(a) = (x - a)g(x, a), \tag{19}$$

where the coefficients of the polynomial $g(x, a)$ depend on a. For $x = a$ the polynomial $g(x, a)$ takes the value $f'(a)$ (this was just our definition of the derivative of a polynomial, see formula (13) of Chapter II). By the assumption, $f'(a) \neq 0$, and so $g(a, a) = f'(a) \neq 0$. Denote by ε an arbitrary number, smaller than the distance from a to the nearest root of the polynomial $g(x, a)$ (here, a is fixed and x is the unknown), so that the polynomial $g(x, a)$ does not vanish on the segment $[a - \varepsilon, a + \varepsilon]$. Then it preserves the same sign on this segment as it has for $x = a$: if it acquired two values of opposite signs, then by Bolzano's theorem it would vanish somewhere inside the segment, which would contradict the choice of the number ε. This contains in fact the assertion of Theorem 6. Let, for example, $f'(a) > 0$. Then $g(a,a) = f'(a) > 0$, too, and according to what was said, $g(x,a) > 0$ for $a-\varepsilon < x < a+\varepsilon$. The other factor $x-a$ in formula (19) also behaves in the known way: $x-a < 0$ for $a-\varepsilon < x < a$ and $x-a > 0$ for $a < x < a+\varepsilon$. Multiplying, we obtain from formula (19) that $f(x)-f(a) < 0$ for $a-\varepsilon < x < a+\varepsilon$ and $f(x)-f(a) > 0$ for $a < x < a+\varepsilon$. This is really the assertion of the theorem. The case $f'(a) < 0$ is treated completely analogously.

The theorem we have just proved has an interesting corollary.

Theorem 7. (Rolle's theorem) *Between two adjacent roots of a polynomial, not having multiple roots, there is always a root of its derivative.*

We assume that our polynomial does not have multiple roots only to make argument shorter. Anyway, this will be the only case that we shall need later.

$f'(\alpha) > 0$, $f'(\beta) > 0$ – impossible

Fig. 6

$f'(a) > 0$, $f'(\beta) < 0$ – possible

Fig. 7

Let α and β, $\alpha < \beta$, be two adjacent roots of the polynomial $f(x)$, so it has no roots lying between them. Since we have assumed that the polynomial has no multiple roots, α and β are not multiple roots and by Theorem 5 of Chapter II, $f'(\alpha) \neq 0$, $f'(\beta) \neq 0$. Let, for example, $f'(\alpha) > 0$. Let us prove that then $f'(\beta) < 0$. Really, if $f'(\beta) > 0$, then by the preceding theorem we would have $f(x) > f(\alpha) = 0$ for $\alpha+\varepsilon > x > \alpha$ and $f(y) < f(\beta) = 0$ for $\beta-\varepsilon < y < \beta$. Then, for arbitrary x satisfying $\alpha+\varepsilon > x > \alpha$ and for arbitrary y, satisfying $\beta-\varepsilon < y < \beta$, we would have $f(x) > 0$ and $f(y) < 0$. Then Bolzano's theorem would imply that the polynomial f had a root lying between x and y, i.e., in the segment $[\alpha,\beta]$. But this would contradict the fact that α and β, as we assumed, were adjacent roots of the polynomial $f(x)$. We see that there remains the only possibility that $f'(\beta) < 0$, but then by Bolzano's theorem the polynomial $f'(x)$ has a root between α and β. On Figs. 6 and 7 an impossible and a possible case of signs for $f'(\beta)$ (if $f'(\alpha) > 0$) are demonstrated. The case when $f'(\alpha) < 0$ can be considered literally in the same way.

At the end of this Section we shall show that the theorems we have proved are already sufficient to solve completely the question about the number of roots for a polynomial of the third degree. In Section 3 of Chapter II we saw that each equation of the third degree can be replaced by an equivalent equation of the form $x^3+ax+b=0$. We shall investigate such a form in the sequel.

First of all let us solve the question about multiple roots. We proved in section 2 of Chapter II that multiple roots of a polynomial are in fact joint roots of the polynomial and its derivative. According to formula (15) of Section II, for the polynomial $f(x)=x^3+ax+b$ the derivative is equal to $f'(x)=3x^2+a$. If $a>0$, then the derivative has no roots and this means that the polynomial $f(x)$ has no multiple roots. If $a<0$, then denote by δ the positive root of the polynomial $3x^2+a$ (i.e., $\delta=\sqrt{-a/3}$). Then the polynomial $f(x)$ can have as a multiple root only one of the numbers δ or $-\delta$. Since the polynomial $f(x)$ can be written in the form $f(x)=(x^2+a)x+b$ and for $x=\pm\delta$, $x^2=-a/3$ and $x^2+a=2a/3$, then the condition that $f(x)$ has a multiple root takes the form $\pm\delta\frac{2a}{3}=-b$, i.e., $\delta^2\frac{4a}{9}=b^2$, and since $\delta^2=-a/3$, the condition becomes $-\frac{4a^3}{27}=b^2$, i.e., $4a^3+27b^2=0$. If this condition is satisfied, then the polynomial has a multiple root α and may be represented in the form $f(x)=(x-\alpha)^2g(x)$. Here the polynomial $g(x)$ has to be of the first degree which means that it has a single root β. Thus, the polynomial $f(x)$ has two roots equal to α, and one root equal to β.

Consider now the remaining case when the polynomial $f(x)$ does not have multiple roots, i.e., $4a^3+27b^2\neq 0$. According to Corollary 2 of Theorem 5, the polynomial $f(x)$ has at least one root α. If it has another root β, then it must be divisible by $(x-\alpha)(x-\beta)$, i.e., it has the form $f(x)=(x-\alpha)(x-\beta)g(x)$, where $g(x)$ is a polynomial of the first degree and therefore it has a root γ. In such a way, the polynomial $f(x)$ has three roots: α, β and γ. It cannot have more than three roots. We conclude that only two things can happen: either the polynomial $f(x)$ has 1 root or the polynomial $f(x)$ has 3 roots. Our problem is to find out which of the cases takes place (for given coefficients a and b).

Fig. 8

Suppose that the polynomial $f(x)$ has three roots: α, β and γ, where $\alpha<\beta<\gamma$. This means that the polynomial does not have roots smaller than α and larger than γ. But according to Corollary 1 of Theorem 5 there exists a number N such that for x large enough (more precisely, for $x \geqslant N$), the values of the polynomial have the same sign as the values of the leading term x^3—i.e., they are positive, and for $x \leqslant -N$ they are negative, for the same reason. Hence, for $x < \alpha$ it is always $f(x) < 0$, and for $x > \gamma$ it is always $f(x) > 0$ (Fig. 8).

Since we have $f(x) < 0$ for $\alpha - \varepsilon < x < \alpha$ and arbitrary $\varepsilon > 0$, according to Theorem 6, $f'(\alpha) > 0$ and so $f(x) > 0$ for $\alpha < x < \alpha + \varepsilon$. Since $f(x)$ has no roots between α and β, by Bolzano's theorem its values are of the fixed sign, so $f(x) > 0$ for $\alpha < x < \beta$. Analogously, we obtain that $f(x) < 0$ for $\beta < x < \gamma$. According to Theorem 7, between the roots α and β, and also between the roots β and γ, there is a root of the derivative $f'(x)$ of the polynomial $f(x)$. Since $f'(x) = 3x^2 + a$, for $a > 0$ the derivative has no roots and such a case (existence of three roots of the polynomial $f(x)$) is impossible. For $a = 0$, $f(x) = x^3 + b$. As we have seen earlier, such a polynomial has only one root. Finally, if $a < 0$, the derivative $f'(x) = 3x^2 + a$ has two roots: $\delta > 0$ and $-\delta < 0$ (here, $\delta = \sqrt{-a/3}$). Obviously, $\alpha < -\delta < \beta < \delta < \gamma$.

Since the polynomial takes positive values on the interval (α, β), and negative values on the interval (β, γ), we have

$$f(-\delta) > 0, \qquad f(\delta) < 0 \tag{20}$$

(under the preposition that the polynomial $f(x)$ has three roots).

Conversely, if conditions (20) are satisfied, then by Bolzano's theorem the polynomial $f(x)$ has a root lying between $-\delta$ and δ. Denote this root by β. Besides, according to Corollary 1 of Theorem 5, for x sufficiently large, the polynomial takes positive values, and for x sufficiently small it takes negative values. Bolzano's theorem implies then that the polynomial has a root smaller than $-\delta$, and also a root greater than δ. Denote these roots by α and γ, respectively. Thus, conditions (20) imply that the polynomial has 3 roots: α, β and γ. In other words, conditions (20) are *necessary and sufficient* for the polynomial $f(x)$ to have 3 roots. In all other cases it has 1 root.

The assertions we have just proved solve our problem. We will only transform conditions (20) into a simpler form. Since $f(x) = (x^2 + a)x + b$ and $3\delta^2 + a = 0$, $\delta^2 = -a/3$, we have $f(\pm\delta) = (\delta^2 + a)(\pm\delta) + b = \pm\delta\frac{2a}{3} + b$ and so conditions (20) acquire the form

$$-\frac{2a}{3}\,\delta + b > 0, \qquad \frac{2a}{3}\,\delta + b < 0,$$

i.e., $\dfrac{2a}{3}\,\delta < b < -\dfrac{2a}{3}\,\delta$. These inequalities are equivalent to just one: $b^2 < \dfrac{4a^2}{3^2}\,\delta^2$. Since $\dfrac{4a^2}{3^2}\,\delta^2 = -\dfrac{4a^3}{27b^2}$, conditions (20) are equivalent to the inequality $4a^3 + 27b^2 < 0$. This is in fact the final answer: if $4a^3 + 27b^2 < 0$, then the polynomial $x^3 + ax + b$ has 3 roots, if $4a^3 + 27b^2 = 0$, it has two equal roots and one other root, and if $4a^3 + 27b^2 > 0$, then it has only 1 root.

Clearly, all that has been said applies only to a polynomial of the third degree. For polynomials of arbitrary degrees analogous investigations can be done, but arguments are a bit more complicated, so we shall leave them for the Appendix.

Problems

1. We proved at the end of Chapter I that the polynomial $x^3 - 7x^2 + 14x - 7$ has no rational roots, so its roots—if they exist—are irrational numbers. Determine the number of roots of this polynomial, their signs and also, for each of the roots, two consecutive integers such that this root is lying between them.

2. Prove that the polynomial $x^4 + ax + b$ either has no roots, or it has two roots and find conditions (on coefficients a and b) such that the first or the latter case takes place.

3. Prove that the number of roots of a polynomial of even degree is even and of odd degree is odd.

4. Prove that the polynomial $x^n + ax + b$, for n even, has 0 or 2 roots, and for n odd—1 or 3. Determine conditions (on coefficients a and b) such that the first or the latter case takes place.

5. Determine the number of roots of the polynomial $x^n + ax^{n-1} + b$ (depending on n, a and b).

6. Prove that each polynomial $f(x)$ takes arbitrarily large values (by absolute value), for sufficiently large values of x (by absolute value).

7. Prove that as a bound of roots N the number $\dfrac{M}{|a_n|} + 1$ can be taken, where M is the largest of the numbers $|a_0|, \ldots, |a_{n-1}|$. *Hint.* Use the inequality $|a_0 + \cdots + a_{n-1}z^{n-1}| \leqslant M(1 + |z| + \cdots + |z|^n)$.

8. Prove that the polynomial $f(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + a_nx^n$, where $a_n > 0$, $a_i \leqslant 0$ for $i = 1, \ldots, n-1$, $a_0 < 0$, has exactly one positive root. *Hint.* Write $f(x)$ in the form $a_nx^n\left(1 + \dfrac{a_{n-1}}{a_nx} + \cdots + \dfrac{a_0}{a_nx^n}\right)$ and find whether the expressions $\dfrac{a_{n-k}}{a_nx^n}$ increase or decrease when x increases, remaining positive.

9. Let a polynomial $f(x)$ have all the coefficients at even powers of x equal to 0, and all the coefficients at odd powers positive. Prove that it has a unique root.

APPENDIX

Sturm's Theorem

We shall present now a method allowing to determine for each polynomial $f(x)$ the number of its roots lying in a given segment $[a, b]$.

The idea of the method is based on the fact that, although for a single polynomial $f(x)$ there is no simple method which could connect its properties with some properties of polynomials with smaller degree, for a pair of polynomials $f(x)$, $g(x)$ such a method is well known: it consists of divison with remainder of the polynomial $f(x)$ by $g(x)$: $f(x) = g(x)q(x) + r(x)$, and passing from the pair of polynomials (f, g) to the pair of polynomials (g, r). Repeating this process leads us to the algorithm of Euclid for finding the greatest common divisor of polynomials f and g.

For example, the question of the existence of common roots of polynomials f and g can be reduced to the question of the existence of common roots of the polynomials of smaller degree g and r and, as a result, to the question of the existence of roots of the polynomial of smaller degree $\mathrm{g.\,c.\,d.}(f, g)$. The method can be applied to the case of the pair of a polynomial and its derivative and then we obtain the answer to the question of the existence of multiple roots of the polynomial. That is how we proceeded in Chapter II, and we shall also proceed like that now: we shall first consider a certain property of roots of the pair of polynomials (f, g), which can be treated using division with remainder. Applying then this property to the pair consisting of a polynomial and its derivative, we shall find the answer to our question.

Let us start with a simple observation, related to a single polynomial $F(x)$. Let $x = \alpha$ be its root and let this root have the multiplicity k. Then we can write down (by the definition of the multiplicity of roots, given in Section 2 of Chapter II)

$$F(x) = (x - \alpha)^k G(x), \tag{1}$$

where $G(\alpha) \neq 0$. Thus, if a number ε is smaller than the distance from α to the nearest root of the polynomial $G(x)$, then $G(x)$ takes the values of the same sign in the segment $[\alpha - \varepsilon, \alpha + \varepsilon]$. Really, if for any two numbers x and y lying in this segment the polynomial G had values $G(x)$ and $G(y)$ of opposite signs, then, by Bolzano's theorem, there would exist a root of the polynomial between x and y. But this would contradict the way how ε had been chosen—that there had been no root of the polynomial G lying in the segment $[\alpha - \varepsilon, \alpha + \varepsilon]$. In particular, all the values of the polynomial $G(x)$ for x in the segment $[\alpha - \varepsilon, \alpha + \varepsilon]$ have the same sign as $G(\alpha)$. Formula (1) implies now that if multiplicity k is even, then the values of the polynomial $F(x)$ for x lying in the segment $[\alpha - \varepsilon, \alpha + \varepsilon]$ have the same sign as $G(\alpha)$. The graph could be situated as in Fig. 1.

$G(\alpha) > 0$ $G(\alpha) < 0$

Fig. 1

If, on the other hand, the multiplicity k is odd, then for $G(\alpha) > 0$ we have $F(x) < 0$ for $\alpha - \varepsilon \leqslant x < \alpha$ and $F(x) > 0$ for $\alpha < x \leqslant \alpha + \varepsilon$, and for $G(\alpha) < 0$—the opposite: $F(x) > 0$ for $\alpha - \varepsilon \leqslant x < \alpha$ and $F(x) < 0$ for $\alpha < x \leqslant \alpha + \varepsilon$. In the former case (i.e., for $G(\alpha) > 0$) α is a *root with increasing*, and in the latter (for $G(\alpha) < 0$)—*root with decreasing*. Possible graphs of the polynomial $F(x)$ in both cases are displayed in Fig. 2.

$G(\alpha) > 0$ $G(\alpha) < 0$

Fig. 2

Definition. Let $F(x)$ be a polynomial having as roots neither a nor b. *Characteristics of the polynomial $F(x)$ on the segment* $[a, b]$ is the difference between the number of its roots with increasing and the roots with decreasing, lying in that segment. Here, roots having even multiplicity are not counted. The characteristics is denoted by $[F(x)]_a^b$. For example, the polynomial represented in Fig. 3 has 3 roots with increasing and 2 roots with decreasing, so we have $[F]_a^b = 1$.

Fig. 3

Since after each root with increasing there must follow a root with decreasing (roots with even multiplicity do not count), the characteristics is determined by the signs of numbers $F(a)$ and $F(b)$, namely:

$[F(x)]_a^b = 0$	if $F(a)$ and $F(b)$ are of the same sign
$[F(x)]_a^b = 1$	if $F(a) < 0$, $F(b) > 0$
$[F(x)]_a^b = -1$	if $F(a) > 0$, $F(b) < 0$

Table 1

Thus, the characteristics of the polynomial $F(x)$ on the given segment is determined by its signs at the endpoints of the segment and so it can be evaluated easily, although by the definition it is connected with its roots which are ususally hard to find.

Our situation can be visually demonstrated as if a passenger is travelling, crossing several times the border between two states, say France and Germany. What is the difference between the number of crossings the border from France to Germany and from Germany to France? Obviously, it is equal to 0 if the passenger started and finished his travel in the same state; it is equal to 1 if he started in France and finished in Germany and to -1 if he started in Germany and finished in France. His itinerary can be demonstrated as a line similar to the graph in Fig. 3, where France is the area below the x-axis and Germany is above.

Consider now two polynomials, f and g, and assume that, first of all, they have no common roots, and, secondly, that the former (i.e., f) does not vanish at $x = a$, nor at $x = b$. *The characteristics of the polynomial f with respect to the polynomial g on the segment* $[a, b]$ is the difference between the number of roots of the polynomial f contained in the segment $[a, b]$ and being roots with increasing of the polynomial fg and the number of its roots being roots with decreasing for fg. The characteristics is denoted by $(f, g)_a^b$.

The main example, which was the reason to introduce this notion is given by the following proposition.

Theorem 1. *If a polynomial $f(x)$ has no multiple roots and neither it nor its derivative vanishes at the endpoints a and b of the segment $[a, b]$, then the characteristics $(f, f')_a^b$ is equal to the number of roots of the polynomial f contained in the segment $[a, b]$.*

The theorem is an easy consequence of Corollary of Theorem 4, Section 3. We simply state that all the roots of the polynomial $f(x)$ are roots with increasing of the polynomial ff'. Really, according to Theorem 5 of Chapter II, the polynomials f and f' have no common roots. If α is a root of the polynomial $f(x)$ with $f'(\alpha) > 0$, then according to Theorem 4 of Section 3 α is a root with increasing for $f(x)$, and so also for $f(x)f'(x)$, since $f'(x) > 0$ in a neighbourhood of α. If, on the other hand, $f'(\alpha) < 0$, then α is a root with decreasing for $f(x)$, and so again a root with increasing for $f(x)f'(x)$, since $f'(x) < 0$ in a neighbourhood of α.

The characteristics $(f, g)_a^b$ is in fact an expression which can be evaluated using division with remainder. Note first the following simple properties:

a) $(f, -g) = -(f, g)$.

This is obvious since when multiplying the polynomial g by -1, the roots with increasing and the roots with decreasing of the polynomial fg interchange.

b) If $g(a) \neq 0$ and $g(b) \neq 0$, then $(f, g)_a^b + (g, f)_a^b = [fg]_a^b$.

This is also obvious since, by the assumption, the polynomials f and g have no common roots. Hence, the roots of the polynomial fg split into the roots of the polynomial f and those of the polynomial g. The number of roots with increasing (and similarly for roots with decreasing) of the polynomial fg is equal to the sum of the numbers of such roots of the polynomial f and of the polynomial g, which gives us the equality b).

c) If polynomials g and h take the same values at the roots of a polynomial f (i.e., if $g(\alpha) = h(\alpha)$ whenever $f(\alpha) = 0$), then

$$(f, g)_a^b = (f, h)_a^b.$$

Really, if $g(\alpha) = h(\alpha)$, then a root α of the polynomial $f(x)$ is at the same time a root with increasing (decreasing) for the polynomials fg and fh.

d) If a polynomial f is divisible by a polynomial g, then

$$(f, g)_a^b = [fg]_a^b.$$

Really, the polynomial g has no roots, since its roots would be common roots for the polynomials f and g. Therefore, $(g, f)_a^b = 0$ and from the property b) it follows that $(f, g)_a^b = [fg]_a^b$.

We shall describe now the process of evaluating the characteristics $(f, g)_a^b$. Divide f by g with remainder:

$$f = gq + r. \tag{2}$$

According to property b), we have $(f, g)_a^b = -(g, f)_a^b + [fg]_a^b$. On the other hand, it follows from relation (2) that $f(\alpha) = r(\alpha)$ whenever $g(\alpha) = 0$. Hence, by property c) we obtain that $(f, g)_a^b = (g, r)_a^b$. The obtained equalities together show that

$$(f, g)_a^b = -(g, r)_a^b + [fg]_a^b. \tag{3}$$

As a matter of fact, relation (3) solves our problem, since it reduces the evaluation of the characteristics $(f, g)_a^b$ to the evaluation of the characteristics $(g, r)_a^b$ for the polynomials g and r of smaller degree, because the expression $[fg]_a^b$ is determined by the values of the polynomials f and g at the endpoints a and b of the segment $[a, b]$ (see Table 1).

Our process of passing from the pair (f, g) to a pair of polynomials with smaller degree is the same as in the process of determining the greatest common divisor of the polynomials f and g. In such a case the characteristics is determined by property d).

We intend to improve our result in two directions. Firstly, we shall present in a unified form the final answer which can be obtained after passing from the pair (f, g) to (g, r) and then executing all the divisions in the consecutive steps of the Euclid's algorithm. Secondly, our inductive reasoning needs that conditions imposed on the polynomials f and g ($f(a) \neq 0$, $f(b) \neq 0$) are then imposed to the polynomials g, r etc. We shall show how one can get rid of these additional restrictions.

First of all, we shall transform a bit the answer we have obtained (formula (3)). We start with changing the notation. The polynomial f will be denoted by f_1, g by f_2 and $-r$ by f_3. Taking into account condition a) of the characteristics, formula (3) obtains the form

$$(f_1, f_2)_a^b = (f_2, f_3)_a^b + [f_1 f_2]_a^b, \tag{4}$$

and the formula of division with remainder (formula (2)) the form

$$f_1 = f_2 q_1 - f_3$$

(we have denoted here q by q_1). Now it is clear how to apply formula (4), reducing degrees of polynomials considered. Starting from f_1 and f_2 define polynomials f_i by induction:

$$f_{i-1} = f_i q_{i-1} - f_{i+1}, \tag{5}$$

where the degree of f_{i+1} is smaller than the degree of f_i (assuming that f_{i-1} and f_i are already defined). Clearly, f_{i-1} are just those plynomials which appear as remainders in the Euclid's algorithm, only with the changed signs. After several steps we come to a polynomial f_k, differing eventually only by sign with the $\gcd(f_1, f_2)$.

Applying formula (4) to f_2 and f_3 instead to f_1 and f_2, we obtain that $(f_2, f_3)_a^b = (f_3, f_4)_a^b + [f_2 f_3]_a^b$. Substituting this value for $(f_2, f_3)_a^b$ into formula (4), we get

$$(f_1, f_2)_a^b = (f_3, f_4)_a^b + [f_1 f_2]_a^b + [f_2 f_3]_a^b.$$

Repeating this process k times and noting that $[f_k f_{k+1}]_a^b = 0$ as a result we obtain:

$$(f_1, f_2)_a^b = [f_1 f_2]_a^b + [f_2 f_3]_a^b + \cdots + [f_{k-1} f_k]_a^b. \tag{6}$$

However, in order that we have the right to apply formula (4), we have to assume that $f_i(a) \neq 0$, $f_i(b) \neq 0$ for all $i = 1, 2, \ldots, k$.

Consider carefully the expression $[fg]_a^b$ which can be evaluated using Table 1 for $F = fg$. In our case it can be rewritten as

$$[fg]_a^b = \begin{cases} 0, & \text{if } f(a)g(a) > 0 \text{ and } f(b)g(b) > 0, \text{ or } f(a)g(a) < 0 \text{ and } f(b)g(b) < 0, \\ 1, & \text{if } f(a)g(a) < 0 \text{ and } f(b)g(b) > 0, \\ -1, & \text{if } f(a)g(a) > 0 \text{ and } f(b)g(b) < 0. \end{cases}$$

Table 2

If two numbers A and B, distinct from 0, are given, then one says that in the pair (A, B) there exists one change of sign if A and B are of opposite signs, and that there is no change of sign if they are of the same sign. Using this terminology, one can reformulate information of Table 2, denoting by n the number of changes of sign in the pair $(f(a), f(b))$ and by m the number of changes of sign in the pair $(f(b), g(b))$. Table 2 obtains the form:

$[fg]_a^b$	m	n
0	0	0
0	1	1
1	1	0
-1	0	1

We see that in all the cases we have $[fg]_a^b = m - n$.

We shall apply now the last remark to formula (6). Denote by m_i the number of changes of sign in the pair $(f_i(a), f_{i+1}(a))$, and by n_i the number of changes of sign in the pair $(f_i(b), f_{i+1}(b))$. As a consequence of the remark, formula (6) obtains the form

$$(f_1, f_2)_a^b = m_1 - n_1 + m_2 - n_2 + \cdots + m_k - n_k. \tag{7}$$

What is the meaning of the number $m_1 + m_2 + \cdots + m_k$? One has just to write down the numbers $f_1(a)$, $f_2(a)$, … , $f_k(a)$ and find out how many changes of sign are there in this sequence—the number of these changes will be $m_1+m_2+\cdots+m_k$. In general, if a sequnce of numbers A_1, … , A_k, distinct from 0, is given, then by the *number of changes of sign* in this sequence, we shall mean the number of places where numbers of opposite signs stay. For example, in the sequence 1, −1, 2, 1, 3, −2 there are 3 changes of sign. We can say that $m_1+m_2+\cdots+m_k$ is the number of changes of sign in the sequence $f_1(a)$, $f_2(a)$, … , $f_k(a)$, and that $n_1+n_2+\cdots+n_k$ is the number of changes of sign in the sequence $f_1(b)$, $f_2(b)$, … , $f_k(b)$. Formula (7) can be interpreted now in the following way:

Theorem 2. *If none of the terms f_1, … , f_k of Sturm's sequence of polynomials f_1, f_2 vanishes, either in a, or in b, and the polynomials f_1, f_2 have no common roots, then the characteristics $(f,g)_a^b$ is equal to the difference between the numbers of changes of sign in the sequences of values of polynomials in Sturm's sequence at the points a and b.*

We have now to get rid of the restrictions $f_i(a) \neq 0$, $f_i(b) \neq 0$ for $i = 1, \dots, k$, which can be uncomfortable in applications: we shall assume just that $f_1(a) \neq 0$ and $f_1(b) \neq 0$. In order to do that we have to generalize a bit the notion of the number of changes of sign. If some of the terms in the sequence A_1, … , A_k are equal to 0, then the number of changes of sign in it is defined as the number of changes of sign in the sequence which is obtained by deleting all the zeros in the given sequence. For example, deleting zeros in the sequence 1, 0, 2, −1, 0, 3, 1, the sequence 1, 2, −1, 3, 1 is obtained, and the latter has two changes of sign. Hence, the given sequence has two changes of sign, by definition.

Denote now by ε the distance from a to the nearest root (distinct from a) of any of the polynomials $f_i(x)$. Thus, $f_i(x) \neq 0$ for $a < x < a + \varepsilon$. Choose any such value a', $a < a' < a + \varepsilon$. A value b' is chosen analogously. Let us state a lemma.

Lemma. *The number of changes of sign in the sequence $f_1(a)$, … , $f_k(a)$ is equal to the number of changes of sign in the sequence $f_1(a')$, … , $f_k(a')$. The same is true when a and a' are replaced by b and b'.*

First of all, let us show that the Lemma can really help us to extend Theorem 2 to arbitrary polynomials f_1, f_2 with the only conditions that $f_1(a) \neq 0$, $f_1(b) \neq 0$ and that f_1 and f_2 have no common roots.

Really, by the assumption, the polynomial f_1 has no roots in the segments $[a, a']$ and $[b', b]$. Hence, all of its roots contained in the segment $[a, b]$, are already contained in the segment $[a', b']$. Therefore, $(f_1, f_2)_a^b = (f_1, f_2)_{a'}^{b'}$. Theorem 2 can now be applied to the characteristics $(f_1, f_2)_{a'}^{b'}$. The number of changes of sign in the sequence $f_1(a'), \dots, f_k(a')$, as well as in the sequence $f_1(b'), \dots, f_k(b')$, is determined by the Lemma. Thus, we obtain the wanted result:

Theorem 3. *If polynomials f_1 and f_2 have no common roots, $f_1(a) \neq 0$ and $f_1(b) \neq 0$, then the characteristics $(f_1, f_2)_a^b$ is equal to the difference between the numbers of changes of sign in the sequence $f_1(a), \dots, f_k(a)$ and in the sequence $f_1(b), \dots, f_k(b)$, where $f_1(x), \dots, f_k(x)$ is Sturm's sequence corresponding to the pair of polynomials f_1, f_2.*

We shall show now that the Lemma is valid. Consider, for example, the value $x = a$. Suppose that $f_i(a) = 0$ for some $i = 1, \dots, k$. By the assumption, $i \neq 1$ since $f_1(a) \neq 0$. Also, $i \neq k$ since the polynomial $f_k(x)$ can differ from $\gcd(f_1, f_2)$ only by sign and so it is a number distinct from 0. Note that then $f_{i-1}(a) \neq 0$ and $f_{i+1}(a) \neq 0$. Really, if we had, for example, $f_i(a) = 0$, $f_{i+1}(a) = 0$, then it would follow from formula (5) that $f_{i-1}(a) = 0$. In exactly the same way, this would imply that $f_{i-2}(a) = 0$ etc., and finally $f_1(a) = 0$, which would contradict the original assumption. But we can say even more—not only that the numbers $f_{i-1}(a)$ and $f_{i+1}(a)$ are distinct from 0, but they have opposite signs—it follows immediately by substituting $x = a$ into equality (5) and taking into account the assumption that $f_i(a) = 0$.

Compare now the sequences $f_1(a), \dots, f_k(a)$ and $f_1(a'), \dots, f_k(a')$. Let $f_i(a) = 0$. Then, as we have seen, $f_{i-1}(a) \neq 0$ and $f_{i+1}(a) \neq 0$, and $f_{i-1}(a)$ and $f_{i+1}(a)$ have opposite signs. But then $f_{i-1}(a') \neq 0$ and $f_{i+1}(a') \neq 0$, and $f_{i-1}(a')$ has the same sign as $f_{i-1}(a)$, while $f_{i+1}(a')$ has the same sign as $f_{i+1}(a)$. This follows from the fact that the polynomials f_{i-1} and f_{i+1} have no roots in the segment $[a, a']$, and so (by Bolzano's theorem) they can have no values of opposite signs. Write down the respective parts of our sequences. Suppose that $f_{i-1}(a) > 0$. Then we obtain the following table:

	$f_{i-1}(x)$	$f_i(x)$	$f_{i+1}(x)$
$x = a$	$+$	0	$-$
$x = a'$	$+$	?	$-$

The characteristics $(f_1, f_2)_{a'}^{b'}$ depends on the number of changes of sign in the lowest row. But we see that it coincides with the number of changes of sign in the row above it—whatever the unkown sign, denoted by ?, is, there will be exactly one change of sign in each of the rows. The case when $f_{i-1}(a) < 0$ can be treated exactly in the same way. The Lemma is proved.

Combining Theorem 3 with Theorem 1 we obtain the basic result:

Theorem 4. (Sturm's Theorem) *If a polynomial $f(x)$ has no multiple roots and does not vanish for $x = a$ and $x = b$, then the number of its roots in the segment $[a, b]$ is equal to the difference between the number of changes of sign of the values of polynomials in the Sturm's sequence, formed for the polynomials $f(x)$ and $f'(x)$ at $x = a$ and $x = b$.*

One has only to note that the lack of multiple roots of the polynomial $f(x)$ is equivalent to the lack of common roots of the polynomials $f(x)$ and $f'(x)$—this is just the assertion of Theorem 5 of Chapter II. Therefore we can apply Theorem 1 to the polynomial $f(x)$ and then Theorem 3 to the pair of polynomials $f(x)$ and $f'(x)$.

Sturm's theorem gives a possibility to answer the basic questions about distribution of roots of a polynomial. First of all, using the theorem, the number of roots can be determined. In order to do that, it is enough to remember Theorem 3 of Section 3, which indicates a number N such that all the roots of the polynomial lie between $-N$ and N. After that it is sufficient to apply Sturm's theorem to the segment $[-N, N]$. However, it is remarkable that in order to determine the number of roots it is neither necessary to evaluate the number N (using Theorem 3), nor to evaluate the values of polynomials in Sturm's sequence for $x = -N$ and $x = N$. Really, for applying Sturm's theorem it is not necessary to know the values $f_i(\pm N)$ themselves, but only their *signs*. That is why it is sufficient to choose a number N large enough, such that the segment $[-N, N]$ contains not only all the roots of the polynomial $f_1(x)$, but also all the roots of all the polynomials $f_i(x)$ of the Sturm's sequence (i.e., we can choose a respective number N_i for each polynomial $f_i(x)$ and take for N the largest of them). According to Corollary 1 of Theorem 3, Section 3, the sign of the value $f_i(N)$, resp. $f_i(-N)$, coincides with the sign of the leading term of the polynomial $f_i(x)$ for $x = N$, resp. $x = -N$. They are determined by the sign of the leading coefficient of the polynomial $f_i(x)$ and by the parity of its degree. Therefore, there is no need to evaluate N and the values $f_i(N)$ and $f_i(-N)$.

When the number of roots is determined, it is possible to indicate segments, each of which contains exactly one root. In order to do that it is already necessary to evaluate the number N, indicated in Theorem 3 of Section 3. After that the segment $[-N, N]$ is divided into two equal parts and using Sturm's theorem the number of roots in each part is found. Then the same is done with the segments $[-N, 0]$ and $[0, N]$ and the process is continued till each of the segments contains only one root.

If it is known that a segment $[a, b]$ contains exactly one root of the polynomial $f(x)$ and the polynomial has no multiple roots, then the values $f(a)$ and $f(b)$ must be of opposite signs. Really, if the root is equal to α, then, according to Theorem 4 of Section 3, for ε small enough, the values $f(\alpha - \varepsilon)$ and $f(\alpha + \varepsilon)$ have the same sign. But $f(\alpha - \varepsilon)$ and $f(a)$ have to be of the same sign—otherwise the polynomial would have one more root in the segment $[\alpha - \varepsilon, \alpha]$. The same is true for the values $f(\alpha + \varepsilon)$ and $f(b)$. Thus, $f(a)$ has the same sign as $f(\alpha - \varepsilon)$, $f(b)$ the same as $f(\alpha + \varepsilon)$, and $f(\alpha - \varepsilon)$ and $f(\alpha + \varepsilon)$ have opposite signs. Hence, $f(a)$ and $f(b)$ have opposite signs. Knowing that, it is possible to evaluate the root α with arbitrary level of accuracy. It is sufficient to divide the segment $[a, b]$ into two parts by a point c and evaluate $f(c)$. Either $f(a)$ and $f(c)$, or $f(c)$ and $f(b)$ have opposite signs. In the former case α is contained in the segment $[a, c]$, and in the latter—in the segment $[c, b]$. After that we continue the process with the segment containing α until we include α in a segment of arbitrary small length. This means that we have evaluated it with arbitrary level of accuracy.

Consider, for example, the polynomial $f(x) = x^3 + 3x - 1$. Applying the criterion from Section 3, we have to evaluate the expression $4a^3 + 27b^2 = 4 \cdot 27 + 27$. Since it is positive, the polynomial has one root. Applying Theorem 3 of Section 3, we find the value $N = 3$. Therefore, the root is contained between -3 and 3, where $f(-3) < 0$, $f(3) > 0$. Since $f(0) < 0$, the root is contained between 0 and 3. Since $f(1) = 3$ and $f(2) = 13$, the root is contained between 0 and 1. In order to find its first decimal, we have to determine in which of the 10 segments (between 0 and 1/10, 1/10 and 2/10, $\dots$, 9/10 and 1) it lies. Put first $x = 1/2$, then $f(x) = 5/8$. Since $f(0)$ and $f(1/2)$ are of opposite signs, the root is contained between 0 and 1/2. Put now $x = 3/10$. Since $f(\frac{3}{10}) = \frac{27}{1000} + \frac{9}{10} - 1 = \frac{27}{1000} - \frac{1}{10} < 0$, the root is contained between 3/10 and 5/10. Finally, $f(\frac{4}{10}) = \frac{64}{1000} + \frac{12}{10} - 1 > 0$. Hence, the root lies between 3/10 and 4/10 and it has the form $\alpha = 0{,}3\dots$.

Since Sturm's theorem has an elegant formulation and a lot of applications, it became widely known immediately after it had been proved. Jacques Sturm, a French mathematician who had proved it, when teaching about the theorem in his lectures, used to say: "Now I will prove a theorem, the name of which I have the honor to bare".

PROBLEMS

1. Construct Sturm's sequence for the polynomials $f(x)$ and $f'(x)$ if $f(x) = x^2 + ax + b$ or $f(x) = x^3 + ax + b$. Using Sturm's theorem deduce again the results about the numbers of roots of these polynomials, obtained already at the end of Section 3. *Hint.* In the case of $f(x) = x^3 + ax + b$ consider separately different cases of possible signs for a and $D = 4a^3 + 27b^2$.

2. Determine, using Sturm's theorem, the number of roots of the polynomial $x^n + ax + b$, depending on n (more precisely, on its parity), a and b.

3. Find the number of roots of the polynomial $x^5 - 5ax^3 + 5a^2x + 2b$. *Hint.* The answer depends on the sign of the expression $a^5 - b^9$.

4. Let a be a root of the derivative $f'(x)$ of a polynomial $f(x)$. Put $f_1(x) = f(x)$, $f_2(x) = f'(x)/(x-a)$. Let $f(x)$ has no multiple roots, and $f_1(x), \dots , f_k(x)$ is Sturm's sequence for the polynomials $f_1(x)$ and $f_2(x)$. Express the number of roots of the polynomial $f(x)$ in terms of the number of changes of sign in the sequences $f_i(N)$, $f_i(a)$ and $f_i(-N)$, $i = 1, \dots, k$ where N is a sufficiently large number.

5. Let two polynomials f_1 and f_2 be given, with degrees n and $n-1$, respectively, and suppose that in their Sturm's sequence the degree of the polynomial $f_i(x)$ is $n - i + 1$ and its leading coefficient is positive. Prove that the polynomial $f_1(x)$ has n roots. Moreover, each of the polynomials $f_i(x)$ has $n - i + 1$ roots, and between each two adjacent roots of the polynomial $f_i(x)$ there lies a root of the polynomial $f_{i+1}(x)$.

6. Let a polynomial $f(x)$ of degree n has n roots. Prove that in the Sturm's sequence (for the polynomials f and f') each polynomial has the degree which is smaller exactly by 1 than the degree of the previous one, and all the leading coefficients are positive. Prove that these conditions are sufficient in order that a polynomial of degree n has n roots.

I. R. Shafarevich, Russian Academy of Sciences, Moscow, Russia