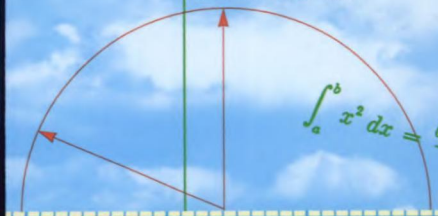
"A lucid representation of the fundamental concepts and methods of the whole field of mathematics."
—Albert Einstein

What is Mathematics?

SECOND EDITION

An Elementary Approach to Ideas and Methods



Richard Courant and Herbert Robbins
Revised by Ian Stewart

WHAT IS

Mathematics?

AN ELEMENTARY APPROACH TO IDEAS AND METHODS

Second Edition

BY

RICHARD COURANT

Late of the
Courant Institute of Mathematical Sciences
New York University

AND

HERBERT ROBBINS

Rutgers University

Revised by

IAN STEWART

Mathematics Institute
University of Warwick

New York *Oxford*
OXFORD UNIVERSITY PRESS
1996

Oxford University Press

Oxford New York
Athens Auckland Bangkok Bogotá Bombay
Buenos Aires Calcutta Cape Town Dar es Salaam
Delhi Florence Hong Kong Istanbul Karachi
Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore
Taipei Tokyo Toronto

and associated companies in
Berlin Ibadan

First published in 1941 by Oxford University Press, Inc.,
198 Madison Avenue, New York, New York 10016

First issued as an Oxford University Press paperback, 1978

First published as a second edition, 1996

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Courant, Richard, 1888–1972.
What is mathematics? : an elementary approach to ideas and methods
/ by Richard Courant and Herbert Robbins.—2nd ed. / revised by Ian Stewart.
p. cm. Includes bibliographical references and index.
ISBN 0-19-510519-2
1. Mathematics. I. Robbins, Herbert. II. Stewart, Ian, 1945– . III. Title.
QA37.2.C69 1996 510—dc20 95-53803

9

Printed in the United States of America
on acid-free paper

CONTENTS

CONTENTS

WHAT IS MATHEMATICS?

Mathematics as an expression of the human mind reflects the active will, the contemplative reason, and the desire for aesthetic perfection. Its basic elements are logic and intuition, analysis and construction, generality and individuality. Though different traditions may emphasize different aspects, it is only the interplay of these antithetic forces and the struggle for their synthesis that constitute the life, usefulness, and supreme value of mathematical science.

Without doubt, all mathematical development has its psychological roots in more or less practical requirements. But once started under the pressure of necessary applications, it inevitably gains momentum in itself and transcends the confines of immediate utility. This trend from applied to theoretical science appears in ancient history as well as in many contributions to modern mathematics by engineers and physicists.

Recorded mathematics begins in the Orient, where, about 2000 B.C., the Babylonians collected a great wealth of material that we would classify today under elementary algebra. Yet as a science in the modern sense mathematics only emerges later, on Greek soil, in the fifth and fourth centuries B.C. The ever-increasing contact between the Orient and the Greeks, beginning at the time of the Persian empire and reaching a climax in the period following Alexander's expeditions, made the Greeks familiar with the achievements of Babylonian mathematics and astronomy. Mathematics was soon subjected to the philosophical discussion that flourished in the Greek city states. Thus Greek thinkers became conscious of the great difficulties inherent in the mathematical concepts of continuity, motion, and infinity, and in the problem of measuring arbitrary quantities by given units. In an admirable effort the challenge was met, and the result, Eudoxus' theory of the geometrical continuum, is an achievement that was only paralleled more than two thousand years later by the modern theory of irrational numbers. The deductive-postulational trend in mathematics originated at the time of Eudoxus and was crystallized in Euclid's *Elements*.

However, while the theoretical and postulational tendency of Greek mathematics remains one of its important characteristics and has exercised an enormous influence, it cannot be emphasized too strongly

that application and connection with physical reality played just as important a part in the mathematics of antiquity, and that a manner of presentation less rigid than Euclid's was very often preferred.

It may be that the early discovery of the difficulties connected with "incommensurable" quantities deterred the Greeks from developing the art of numerical reckoning achieved before in the Orient. Instead they forced their way through the thicket of pure axiomatic geometry. Thus one of the strange detours of the history of science began, and perhaps a great opportunity was missed. For almost two thousand years the weight of Greek geometrical tradition retarded the inevitable evolution of the number concept and of algebraic manipulation, which later formed the basis of modern science.

After a period of slow preparation, the revolution in mathematics and science began its vigorous phase in the seventeenth century with analytic geometry and the differential and integral calculus. While Greek geometry retained an important place, the Greek ideal of axiomatic crystallization and systematic deduction disappeared in the seventeenth and eighteenth centuries. Logically precise reasoning, starting from clear definitions and non-contradictory, "evident" axioms, seemed immaterial to the new pioneers of mathematical science. In a veritable orgy of intuitive guesswork, of cogent reasoning interwoven with nonsensical mysticism, with a blind confidence in the superhuman power of formal procedure, they conquered a mathematical world of immense riches. Gradually the ecstasy of progress gave way to a spirit of critical self-control. In the nineteenth century the immanent need for consolidation and the desire for more security in the extension of higher learning that was prompted by the French revolution, inevitably led back to a revision of the foundations of the new mathematics, in particular of the differential and integral calculus and the underlying concept of limit. Thus the nineteenth century not only became a period of new advances, but was also characterized by a successful return to the classical ideal of precision and rigorous proof. In this respect it even surpassed the model of Greek science. Once more the pendulum swung toward the side of logical purity and abstraction. At present we still seem to be in this period, although it is to be hoped that the resulting unfortunate separation between pure mathematics and the vital applications, perhaps inevitable in times of critical revision, will be followed by an era of closer unity. The regained internal strength and, above all, the enormous simplification attained on the basis of clearer comprehension make it

possible today to master the mathematical theory without losing sight of applications. To establish once again an organic union between pure and applied science and a sound balance between abstract generality and colorful individuality may well be the paramount task of mathematics in the immediate future.

This is not the place for a detailed philosophical or psychological analysis of mathematics. Only a few points should be stressed. There seems to be a great danger in the prevailing overemphasis on the deductive-postulational character of mathematics. True, the element of constructive invention, of directing and motivating intuition, is apt to elude a simple philosophical formulation; but it remains the core of any mathematical achievement, even in the most abstract fields. If the crystallized deductive form is the goal, intuition and construction are at least the driving forces. A serious threat to the very life of science is implied in the assertion that mathematics is nothing but a system of conclusions drawn from definitions and postulates that must be consistent but otherwise may be created by the free will of the mathematician. If this description were accurate, mathematics could not attract any intelligent person. It would be a game with definitions, rules, and syllogisms, without motive or goal. The notion that the intellect can create meaningful postulational systems at its whim is a deceptive halftruth. Only under the discipline of responsibility to the organic whole, only guided by intrinsic necessity, can the free mind achieve results of scientific value.

While the contemplative trend of logical analysis does not represent all of mathematics, it has led to a more profound understanding of mathematical facts and their interdependence, and to a clearer comprehension of the essence of mathematical concepts. From it has evolved a modern point of view in mathematics that is typical of a universal scientific attitude.

Whatever our philosophical standpoint may be, for all purposes of scientific observation an object exhausts itself in the totality of possible relations to the perceiving subject or instrument. Of course, mere perception does not constitute knowledge and insight; it must be coordinated and interpreted by reference to some underlying entity, a "thing in itself," which is not an object of direct physical observation, but belongs to metaphysics. Yet for scientific procedure it is important to discard elements of metaphysical character and to consider observable facts always as the ultimate source of notions and constructions. To renounce the goal of comprehending the "thing in itself," of knowing

the "ultimate truth," of unraveling the innermost essence of the world, may be a psychological hardship for naive enthusiasts, but in fact it was one of the most fruitful turns in modern thinking.

Some of the greatest achievements in physics have come as a reward for courageous adherence to the principle of eliminating metaphysics. When Einstein tried to reduce the notion of "simultaneous events occurring at different places" to observable phenomena, when he unmasked as a metaphysical prejudice the belief that this concept must have a scientific meaning in itself, he had found the key to his theory of relativity. When Niels Bohr and his pupils analyzed the fact that any physical observation must be accompanied by an effect of the observing instrument on the observed object, it became clear that the sharp simultaneous fixation of position and velocity of a particle is not possible in the sense of physics. The far-reaching consequences of this discovery, embodied in the modern theory of quantum mechanics, are now familiar to every physicist. In the nineteenth century the idea prevailed that mechanical forces and motions of particles in space are things in themselves, while electricity, light, and magnetism should be reduced to or "explained" as mechanical phenomena, just as had been done with heat. The "ether" was invented as a hypothetical medium capable of not entirely explained mechanical motions that appear to us as light or electricity. Slowly it was realized that the ether is of necessity unobservable; that it belongs to metaphysics and not to physics. With sorrow in some quarters, with relief in others, the mechanical explanations of light and electricity, and with them the ether, were finally abandoned.

A similar situation, even more accentuated, exists in mathematics. Throughout the ages mathematicians have considered their objects, such as numbers, points, etc., as substantial things in themselves. Since these entities had always defied attempts at an adequate description, it slowly dawned on the mathematicians of the nineteenth century that the question of the meaning of these objects as substantial things does not make sense within mathematics, if at all. The only relevant assertions concerning them do not refer to substantial reality; they state only the interrelations between mathematically "undefined objects" and the rules governing operations with them. What points, lines, numbers "actually" *are* cannot and need not be discussed in mathematical science. What matters and what corresponds to "verifiable" fact is structure and relationship, that two points determine a line, that numbers combine according to certain rules to form other numbers, etc. A clear insight into the necessity of a dissubstantiation of elementary mathematical

concepts has been one of the most important and fruitful results of the modern postulational development.

Fortunately, creative minds forget dogmatic philosophical beliefs whenever adherence to them would impede constructive achievement. For scholars and layman alike it is not philosophy but active experience in mathematics itself that alone can answer the question: What is mathematics?

CHAPTER I

THE NATURAL NUMBERS

INTRODUCTION

Number is the basis of modern mathematics. But what is number? What does it mean to say that $\frac{1}{2} + \frac{1}{2} = 1$, $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, and $(-1)(-1) = 1$? We learn in school the mechanics of handling fractions and negative numbers, but for a real understanding of the number system we must go back to simpler elements. While the Greeks chose the geometrical concepts of point and line as the basis of their mathematics, it has become the modern guiding principle that all mathematical statements should be reducible ultimately to statements about the *natural numbers*, 1, 2, 3, ... "God created the natural numbers; everything else is man's handiwork." In these words Leopold Kronecker (1823-1891) pointed out the safe ground on which the structure of mathematics can be built.

Created by the human mind to count the objects in various assemblages, numbers have no reference to the individual characteristics of the objects counted. The number six is an abstraction from all actual collections containing six things; it does not depend on any specific qualities of these things or on the symbols used. Only at a rather advanced stage of intellectual development does the abstract character of the idea of number become clear. To children, numbers always remain connected with tangible objects such as fingers or beads, and primitive languages display a concrete number sense by providing different sets of number words for different types of objects.

Fortunately, the mathematician as such need not be concerned with the philosophical nature of the transition from collections of concrete objects to the abstract number concept. We shall therefore accept the natural numbers as given, together with the two fundamental operations, addition and multiplication, by which they may be combined.

§1. CALCULATION WITH INTEGERS

1. Laws of Arithmetic

The mathematical theory of the natural numbers or *positive integers* is known as *arithmetic*. It is based on the fact that the addition and

multiplication of integers are governed by certain laws. In order to state these laws in full generality we cannot use symbols like 1, 2, 3 which refer to specific integers. The statement

$$1 + 2 = 2 + 1$$

is only a particular instance of the general law that the sum of two integers is the same regardless of the order in which they are considered. Hence, when we wish to express the fact that a certain relation between integers is valid irrespective of the values of the particular integers involved, we shall denote integers symbolically by letters $a, b, c, \cdots$. With this agreement we may state five fundamental laws of arithmetic with which the reader is familiar:

$$1)\ a + b = b + a, \qquad 2)\ ab = ba,$$

$$3)\ a + (b + c) = (a + b) + c, \qquad 4)\ a(bc) = (ab)c,$$

$$5)\ a(b + c) = ab + ac.$$

The first two of these, the *commutative* laws of addition and multiplication, state that one may interchange the order of the elements involved in addition or multiplication. The third, the *associative* law of addition, states that addition of three numbers gives the same result whether we add to the first the sum of the second and third, or to the third the sum of the first and second. The fourth is the associative law of multiplication. The last, the *distributive* law, expresses the fact that to multiply a sum by an integer we may multiply each term of the sum by this integer and then add the products.

These laws of arithmetic are very simple, and may seem obvious. But they might not be applicable to entities other than integers. If a and b are symbols not for integers but for chemical substances, and if "addition" is used in a colloquial sense, it is evident that the commutative law will not always hold. For example, if sulphuric acid is added to water, a dilute solution is obtained, while the addition of water to pure sulphuric acid may result in disaster to the experimenter. Similar illustrations will show that in this type of chemical "arithmetic" the associative and distributive laws of addition may also fail. Thus one can imagine types of arithmetic in which one or more of the laws 1)–5) do not hold. Such systems have actually been studied in modern mathematics.

A concrete model for the abstract concept of integer will indicate the intuitive basis on which the laws 1)–5) rest. Instead of using the usual number symbols 1, 2, 3, etc., let us denote the integer that gives the

number of objects in a given collection (say the collection of apples on a particular tree) by a set of dots placed in a rectangular box, one dot for each object. By operating with these boxes we may investigate the laws of the arithmetic of integers. To add two integers a and b, we place the corresponding boxes end to end and remove the partition.

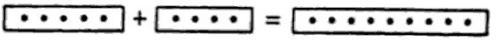


Fig. 1. Addition.

To multiply a and b, we arrange the dots in the two boxes in rows, and form a new box with a rows and b columns of dots. The rules 1)–5)

Fig. 2. Multiplication.

will now be seen to correspond to intuitively obvious properties of these operations with boxes.

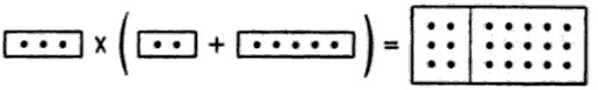


Fig. 3. The Distributive Law.

On the basis of the definition of addition of two integers we may define the relation of *inequality*. Each of the equivalent statements, $a < b$ (read, "a is less than b") and $b > a$ (read, "b is greater than a"), means that box b may be obtained from box a by the addition of a properly chosen third box c, so that $b = a + c$. When this is so we write

$$c = b - a,$$

which defines the operation of *subtraction*.

Fig. 4. Subtraction.

Addition and subtraction are said to be *inverse operations*, since if the addition of the integer d to the integer a is followed by the subtraction of the integer d, the result is the original integer a:

$$(a + d) - d = a.$$

It should be noted that the integer $b - a$ has been defined only when $b > a$. The interpretation of the symbol $b - a$ as a *negative integer* when $b < a$ will be discussed later (p. 54 et seq.).

It is often convenient to use one of the notations, $b \geq a$ (read, "b is greater than or equal to a") or $a \leq b$ (read, "a is less than or equal to b"), to express the denial of the statement, $a > b$. Thus, $2 \geq 2$, and $3 \geq 2$.

We may slightly extend the domain of positive integers, represented by boxes of dots, by introducing the integer *zero*, represented by a completely empty box. If we denote the empty box by the usual symbol 0, then, according to our definition of addition and multiplication,

$$a + 0 = a,$$
$$a \cdot 0 = 0,$$

for every integer a. For $a + 0$ denotes the addition of an empty box to the box a, while $a \cdot 0$ denotes a box with no columns; i.e. an empty box. It is then natural to extend the definition of subtraction by setting

$$a - a = 0$$

for every integer a. These are the characteristic arithmetical properties of zero.

Geometrical models like these boxes of dots, such as the ancient abacus, were widely used for numerical calculations until late in the middle ages, when they were slowly displaced by greatly superior symbolic methods based on the decimal system.

2. The Representation of Integers

We must carefully distinguish between an integer and the symbol, 5, V, $\cdots$, etc., used to represent it. In the decimal system the ten digit symbols, 0, 1, 2, 3, $\cdots$, 9, are used for zero and the first nine positive integers. A larger integer, such as "three hundred and seventy-two," can be expressed in the form

$$300 + 70 + 2 = 3 \cdot 10^2 + 7 \cdot 10 + 2,$$

and is denoted in the decimal system by the symbol 372. Here the important point is that the meaning of the digit symbols 3, 7, 2 depends on their *position* in the units, tens, or hundreds place. With this "positional notation" we can represent any integer by using only the ten digit symbols in various combinations. The general rule is to express an integer in the form illustrated by

$$z = a \cdot 10^3 + b \cdot 10^2 + c \cdot 10 + d,$$

where the digits a, b, c, d are integers from zero to nine. The integer z is then represented by the abbreviated symbol

$$abcd.$$

We note in passing that the coefficients d, c, b, a are the remainders left after successive divisions of z by 10. Thus

$10\overline{)372}$	Remainder
$10\overline{)37}$	2
$10\overline{)3}$	7
0	3

The particular expression given above for z can only represent integers less than ten thousand, since larger integers will require five or more digit symbols. If z is an integer between ten thousand and one hundred thousand, we can express it in the form

$$z = a \cdot 10^4 + b \cdot 10^3 + c \cdot 10^2 + d \cdot 10 + e,$$

and represent it by the symbol *abcde*. A similar statement holds for integers between one hundred thousand and one million, etc. It is very useful to have a way of indicating the result in perfect generality by a single formula. We may do this if we denote the different coefficients, $e, d, c, \cdots$, by the single letter a with different "subscripts," $a_0, a_1, a_2, a_3, \cdots$, and indicate the fact that the powers of ten may be as large as necessary by denoting the highest power, not by 10^3 or 10^4 as in the examples above, but by 10^n, where n is understood to stand for an arbitrary integer. Then the general method for representing an integer z in the decimal system is to express z in the form

$$(1) \qquad z = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \cdots + a_1 \cdot 10 + a_0,$$

and to represent it by the symbol

$$a_n a_{n-1} a_{n-2} \cdots a_1 a_0.$$

As in the special case above, we see that the digits $a_0, a_1, a_2, \cdots, a_n$ are simply the successive remainders when z is divided repeatedly by 10.

In the decimal system the number ten is singled out to serve as a base. The layman may not realize that the selection of ten is not essential, and that any integer greater than one would serve the same purpose. For example, a *septimal* system (base 7) could be used. In such a system, an integer would be expressed as

$$(2) \qquad b_n \cdot 7^n + b_{n-1} \cdot 7^{n-1} + \cdots + b_1 \cdot 7 + b_0,$$

where the b's are digits from zero to six, and denoted by the symbol

$$b_n b_{n-1} \cdots b_1 b_0 .$$

Thus "one hundred and nine" would be denoted in the septimal system by the symbol 214, meaning

$$2 \cdot 7^2 + 1 \cdot 7 + 4.$$

As an exercise the reader may prove that the general rule for passing from the base ten to any other base B is to perform successive divisions of the number z by B; the remainders will be the digits of the number in the system with base B. For example:

$7\overline{)109}$	Remainder
$7\overline{)15}$	4
$7\overline{)2}$	1
0	2

109 (decimal system) = 214 (septimal system).

It is natural to ask whether any particular choice of base would be most desirable. We shall see that too small a base has disadvantages, while a large base requires the learning of many digit symbols, and an extended multiplication table. The choice of twelve as base has been advocated, since twelve is exactly divisible by two, three, four, and six, and, as a result, work involving division and fractions would often be simplified. To write any integer in terms of the base twelve (duodecimal system), we require two new digit symbols for ten and eleven. Let us write α for ten and β for eleven. Then in the duodecimal system "twelve" would be written 10, "twenty-two" would be 1α, "twenty-three" would be 1β, and "one hundred thirty-one" would be $\alpha\beta$.

The invention of positional notation, attributed to the Sumerians or Babylonians and developed by the Hindus, was of enormous significance for civilization. Early systems of numeration were based on a purely additive principle. In the Roman symbolism, for example, one wrote

CXVIII = one hundred + ten + five + one + one + one.

The Egyptian, Hebrew, and Greek systems of numeration were on the same level. One disadvantage of any purely additive notation is that more and more new symbols are needed as numbers get larger. (Of course, early scientists were not troubled by our modern astronomical or atomic magnitudes.) But the chief fault of ancient systems, such as the Roman, was that computation with numbers was so difficult that only the specialist could handle any but the simplest problems. It is quite different with the Hindu positional system now in use. This was introduced into medieval Europe by the merchants of Italy, who learned

it from the Moslems. The positional system has the agreeable property that all numbers, however large or small, can be represented by the use of a small set of different digit symbols (in the decimal system, these are the "Arabic numerals" 0, 1, 2, $\cdots$, 9). Along with this goes the more important advantage of ease of computation. The rules of reckoning with numbers represented in positional notation can be stated in the form of addition and multiplication tables for the digits that can be memorized once and for all. The ancient art of computation, once confined to a few adepts, is now taught in elementary school. There are not many instances where scientific progress has so deeply affected and facilitated everyday life.

3. Computation in Systems Other than the Decimal

The use of ten as a base goes back to the dawn of civilization, and it undoubtedly due to the fact that we have ten fingers on which to count. But the number words of many languages show remnants of the use of other bases, notably twelve and twenty. In English and German the words for 11 and 12 are not constructed on the decimal principle of combining 10 with the digits, as are the "teens," but are linguistically independent of the words for 10. In French the words "vingt" and "quatre-vingt" for 20 and 80 suggest that for some purposes a system with base 20 might have been used. In Danish the word for 70, "halvfirsindstyve," means half-way (from three times) to four times twenty. The Babylonian astronomers had a system of notation that was partly sexagesimal (base 60), and this is believed to account for the customary division of the hour and the angular degree into 60 minutes.

In a system other than the decimal the rules of arithmetic are the same, but one must use different tables for the addition and multiplication of digits. Accustomed to the decimal system and tied to it by the number words of our language, we might at first find this a little confusing. Let us try an example of multiplication in the septimal system. Before proceeding, it is advisable to write down the tables we shall have to use:

Addition

	1	2	3	4	5	6
1	2	3	4	5	6	10
2	3	4	5	6	10	11
3	4	5	6	10	11	12
4	5	6	10	11	12	13
5	6	10	11	12	13	14
6	10	11	12	13	14	15

Multiplication

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	11	13	15
3	3	6	12	15	21	24
4	4	11	15	22	26	33
5	5	13	21	26	34	42
6	6	15	24	33	42	51

Let us now multiply 265 by 24, where these number symbols are written in the septimal system. (In the decimal system this would be equivalent to multiplying 145 by 18.) The rules of multiplication are the same as in the decimal system. We begin by multiplying 5 by 4, which is 26, as the multiplication table shows.

$$\begin{array}{r} 265 \\ 24 \\ \hline 1456 \\ 563 \\ \hline 10416 \end{array}$$

We write down 6 in the units place, "carrying" the 2 to the next place. Then we find $4 \cdot 6 = 33$, and $33 + 2 = 35$. We write down 5, and proceed in this way until everything has been multiplied out. Adding $1{,}456 + 5{,}630$, we get $6 + 0 = 6$ in the units place, $5 + 3 = 11$ in the sevens place. Again we write down 1 and keep 1 for the forty-nines place, where we have $1 + 6 + 4 = 14$. The final result is $265 \cdot 24 = 10{,}416$.

To check this result we may multiply the same numbers in the decimal system. 10,416 (septimal system) may be written in the decimal system by finding the powers of 7 up to the fourth: $7^2 = 49$, $7^3 = 343$, $7^4 = 2{,}401$. Hence $10{,}416 = 2{,}401 + 4 \cdot 49 + 7 + 6$, this evaluation being in the decimal system. Adding these numbers we find that 10,416 in the septimal system is equal to 2,610 in the decimal system. Now we multiply 145 by 18 in the decimal system; the result is 2,610, so the calculations check.

Exercises: 1) Set up the addition and multiplication tables in the duodecimal system and work some examples of the same sort.

2) Express "thirty" and "one hundred and thirty-three" in the systems with the bases 5, 7, 11, 12.

3) What do the symbols 11111 and 21212 mean in these systems?

4) Form the addition and multiplication tables for the bases 5, 11, 13.

From a theoretical point of view, the positional system with the base 2 is singled out as the one with the smallest possible base. The only digits in this *dyadic system* are 0 and 1; every other number z is represented by a row of these symbols. The addition and multiplication tables consist merely of the rules $1 + 1 = 10$ and $1 \cdot 1 = 1$. But the disadvantage of this system is obvious: long expressions are needed to represent small numbers. Thus seventy-nine, which may be expressed as $1 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2 + 1$, is written in the dyadic system as 1,001,111.

As an illustration of the simplicity of multiplication in the dyadic system, we shall multiply seven and five, which are respectively 111 and 101. Remembering that $1 + 1 = 10$ in this system, we have

$$\begin{array}{r} 111 \\ 101 \\ \hline 111 \\ 111 \\ \hline 100011 \end{array} = 2^5 + 2 + 1,$$

which is thirty-five, as it should be.

Gottfried Wilhelm Leibniz (1646–1716), one of the greatest intellects of his time, was fond of the dyadic system. To quote Laplace: "Leibniz saw in his binary arithmetic the image of creation. He imagined that Unity represented God, and zero the void; that the Supreme Being drew all beings from the void, just as unity and zero express all numbers in his system of numeration."

Exercise: Consider the question of representing integers with the base a. In order to name the integers in this system we need words for the digits $0, 1, \cdots, a-1$ and for the various powers of a: $a, a^2, a^3, \cdots$. How many different number words are needed to name all numbers from zero to one thousand, for $a = 2, 3, 4, 5, \cdots, 15$? Which base requires the fewest? (Examples: If $a = 10$, we need ten words for the digits, plus words for 10, 100, and 1000, making a total of 13. For $a = 20$, we need twenty words for the digits, plus words for 20 and 400, making a total of 22. If $a = 100$, we need 100 plus 1.)

*§2. THE INFINITUDE OF THE NUMBER SYSTEM. MATHEMATICAL INDUCTION

1. The Principle of Mathematical Induction

There is no end to the sequence of integers 1, 2, 3, 4, $\cdots$; for after any integer n has been reached we may write the next integer, $n + 1$. We express this property of the sequence of integers by saying that there are *infinitely many* integers. The sequence of integers represents the simplest and most natural example of the mathematical infinite, which plays a dominant rôle in modern mathematics. Everywhere in this book we shall have to deal with collections or "sets" containing infinitely many mathematical objects, like the set of all points on a line or the set of all triangles in a plane. The infinite sequence of integers is the simplest example of an infinite set.

The step by step procedure of passing from n to $n + 1$ which generates the infinite sequence of integers also forms the basis of one of the most fundamental patterns of mathematical reasoning, the principle of

mathematical induction. "Empirical induction" in the natural sciences proceeds from a particular series of observations of a certain phenomenon to the statement of a general law governing all occurrences of this phenomenon. The degree of certainty with which the law is thereby established depends on the number of single observations and confirmations. This sort of inductive reasoning is often entirely convincing; the prediction that the sun will rise tomorrow in the east is as certain as anything can be, but the character of this statement is not the same as that of a theorem proved by strict logical or mathematical reasoning.

In quite a different way *mathematical induction* is used to establish the truth of a mathematical theorem for an infinite sequence of cases, the first, the second, the third, and so on without exception. Let us denote by A a statement that involves an arbitrary integer n. For example, A may be the statement, "The sum of the angles in a convex polygon of $n + 2$ sides is n times 180 degrees." Or A' may be the assertion, "By drawing n lines in a plane we cannot divide the plane into more than 2^n parts." To prove such a theorem for *every* integer n it does not suffice to prove it separately for the first 10 or 100 or even 1000 values of n. This indeed would correspond to the attitude of empirical induction. Instead, we must use a method of strictly mathematical and non-empirical reasoning whose character will be indicated by the following proofs for the special examples A and A'. In the case A, we know that for $n = 1$ the polygon is a triangle, and from elementary geometry the sum of the angles is known to be $1 \cdot 180°$. For a quadrilateral, $n = 2$, we draw a diagonal which divides the quadrilateral into two triangles. This shows immediately that the sum of the angles of the quadrilateral is equal to the sum of the angles in the two triangles, which yields $180° + 180° = 2 \cdot 180°$. Proceeding to the case of a pentagon with 5 edges, $n = 3$, we decompose it into a triangle plus a quadrilateral. Since the latter has the angle sum $2 \cdot 180°$, as we have just proved, and since the triangle has the angle sum 180°, we obtain $3 \cdot 180$ degrees for the 5-gon. Now it is clear that we can proceed indefinitely in the same way, proving the theorem for $n = 4$, then for $n = 5$, and so on. Each statement follows in the same way from the preceding one, so that the general theorem A can be established for all n.

Similarly we can prove the theorem A'. For $n = 1$ it is obviously true, since a single line divides the plane into 2 parts. Now add a second line. Each of the previous parts will be divided into two new parts, unless the new line is parallel to the first. In either case, for

$n = 2$ we have not more than $4 = 2^2$ parts. Now we add a third line. Each of the previous domains will either be cut into two parts or be left untouched. Thus the sum of parts is not greater than $2^2 \cdot 2 = 2^3$. Knowing this to be true, we can prove the next case in the same way, and so on indefinitely.

The essential idea in the preceding arguments is to establish a general theorem A for all values of n by successively proving a sequence of special cases, $A_1, A_2, \cdots$. The possibility of doing this depends on two things: a) There is a general method for showing that *if* any statement A_r is true then the next statement, A_{r+1}, will *also* be true. b) The first statement A_1 is *known* to be true. That these two conditions are sufficient to establish the truth of *all* the statements $A_1, A_2, A_3, \cdots$ is a logical principle which is as fundamental to mathematics as are the classical rules of Aristotelian logic. We formulate it as follows:

Let us suppose that we wish to establish a whole infinite sequence of mathematical propositions

$$A_1, A_2, A_3, \cdots$$

which together constitute the general proposition A. *Suppose that* a) *by some mathematical argument it is shown that if* r *is any integer and* ***if*** *the assertion* A_r *is known to be true* ***then*** *the truth of the assertion* A_{r+1} *will follow, and that* b) *the first proposition* A_1 ***is*** *known to be true. Then all the propositions of the sequence must be true, and* A *is proved.*

We shall not hesitate to accept this, just as we accept the simple rules of ordinary logic, as a basic principle of mathematical reasoning. For we can establish the truth of every statement A_n, starting from the given assertion b) that A_1 is true, and proceeding by repeated use of the assertion a) to establish successively the truth of A_2, A_3, A_4, and so on until we reach the statement A_n. The principle of mathematical induction thus rests on the fact that after any integer r there is a next, $r + 1$, and that any desired integer n may be reached by a finite number of such steps, starting from the integer 1.

Often the principle of mathematical induction is applied without explicit mention, or is simply indicated by a casual "etc." or "and so on." This is especially frequent in elementary instruction. But the explicit use of an inductive argument is indispensable in more subtle proofs. We shall give a few illustrations of a simple but not quite trivial character.

2. The Arithmetical Progression

For every value of n, *the sum* $1 + 2 + 3 + \cdots + n$ *of the first* n *integers is equal to* $\frac{n(n+1)}{2}$. In order to prove this theorem by mathematical induction we must show that for every n the assertion A_n:

$$(1) \qquad 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

is true. a) We observe that if r is an integer and if the statement A_r is known to be true, i.e. if it is known that

$$1 + 2 + 3 + \cdots + r = \frac{r(r+1)}{2},$$

then by adding the number $(r + 1)$ to both sides of this equation we obtain the equation

$$1 + 2 + 3 + \cdots + r + (r+1) = \frac{r(r+1)}{2} + (r+1)$$

$$= \frac{r(r+1) + 2(r+1)}{2} = \frac{(r+1)(r+2)}{2},$$

which is precisely the statement A_{r+1}. b) The statement A_1 is obviously true, since $1 = \frac{1 \cdot 2}{2}$. Hence, by the principle of mathematical induction, the statement A_n is true for every n, as was to be proved.

Ordinarily this is shown by writing the sum $1 + 2 + 3 + \cdots + n$ in two forms:

$$S_n = 1 + 2 + \cdots + (n-1) + n$$

and

$$S_n = n + (n-1) + \cdots + 2 + 1.$$

On adding, we see that each pair of numbers in the same column yields the sum $n + 1$, and, since there are n columns in all, it follows that

$$2S_n = n(n+1),$$

which proves the desired result.

From (1) we may immediately derive the formula for the sum of the first $(n+1)$ terms of any *arithmetical progression*,

$$(2)\quad P_n = a+(a+d)+(a+2d)+\cdots+(a+nd) = \frac{(n+1)(2a+nd)}{2}.$$

For

$$P_n = (n+1)a+(1+2+\cdots+n)d = (n+1)a+\frac{n(n+1)d}{2}$$

$$= \frac{2(n+1)a+n(n+1)d}{2} = \frac{(n+1)(2a+nd)}{2}.$$

For the case $a=0$, $d=1$, this is equivalent to (1).

3. The Geometrical Progression

One may treat the general geometrical progression in a similar way. We shall prove that for every value of n

$$(3)\qquad G_n = a+aq+aq^2+\cdots+aq^n = a\frac{1-q^{n+1}}{1-q}.$$

(We suppose that $q \neq 1$, since otherwise the right side of (3) has no meaning.)

Certainly this assertion is true for $n=1$, for then it states that

$$G_1 = a+aq = \frac{a(1-q^2)}{1-q} = \frac{a(1+q)(1-q)}{(1-q)} = a(1+q).$$

And *if* we assume that

$$G_r = a+aq+\cdots+aq^r = a\frac{1-q^{r+1}}{1-q},$$

then we find as a consequence that

$$G_{r+1} = (a+aq+\cdots+aq^r)+aq^{r+1} = G_r+aq^{r+1} = a\frac{1-q^{r+1}}{1-q}+aq^{r+1}$$

$$= a\frac{(1-q^{r+1})+q^{r+1}(1-q)}{1-q} = a\frac{1-q^{r+1}+q^{r+1}-q^{r+2}}{1-q} = a\frac{1-q^{r+2}}{1-q}.$$

But this is precisely the assertion (3) for the case $n=r+1$. This completes the proof.

In elementary textbooks the usual proof proceeds as follows. Set

$$G_n = a+aq+\cdots+aq^n,$$

and multiply both sides of this equation by q, obtaining

$$qG_n = aq+aq^2+\cdots+aq^{n+1}.$$

Now subtract corresponding sides of this equation from the preceding equation, obtaining

$$G_n - qG_n = a - aq^{n+1},$$
$$(1 - q)G_n = a(1 - q^{n+1}),$$
$$G_n = a\frac{1 - q^{n+1}}{1 - q}.$$

4. The Sum of the First n Squares

A further interesting application of the principle of mathematical induction refers to the sum of the first n squares. By direct trial one finds that, at least for small values of n,

$$(4) \qquad 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

and one might *guess* that this remarkable formula is valid for *all integers* n. To *prove* this, we shall again use the principle of mathematical induction. We begin by observing that *if* the assertion A_n, which in this case is the equation (4), is true for the case $n = r$, so that

$$1^2 + 2^2 + 3^2 + \cdots + r^2 = \frac{r(r+1)(2r+1)}{6},$$

then on adding $(r+1)^2$ to both sides of this equation we obtain

$$1^2 + 2^2 + 3^2 + \cdots + r^2 + (r+1)^2 = \frac{r(r+1)(2r+1)}{6} + (r+1)^2$$
$$= \frac{r(r+1)(2r+1) + 6(r+1)^2}{6} = \frac{(r+1)[r(2r+1) + 6(r+1)]}{6}$$
$$= \frac{(r+1)(2r^2 + 7r + 6)}{6} = \frac{(r+1)(r+2)(2r+3)}{6},$$

which is precisely the assertion A_{r+1} in this case, since it is obtained by substituting $r + 1$ for n in (4). To complete the proof we need only remark that the assertion A_1, in this case the equation

$$1^2 = \frac{1(1+1)(2+1)}{6},$$

is obviously true. Hence the equation (4) is true for every n.

Formulas of a similar sort may be found for higher powers of the integers, $1^k + 2^k + 3^k + \cdots + n^k$, where k is any positive integer As an exercise, the reader may prove by mathematical induction that

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \left[\frac{n(n+1)}{2}\right]^2. \tag{5}$$

It should be remarked that although the principle of mathematical induction suffices to *prove* the formula (5) once this formula has been written down, the proof gives no indication of how this formula was arrived at in the first place; why precisely the expression $[n(n+1)/2]^2$ should be guessed as an expression for the sum of the first n cubes, rather than $[n(n+1)/3]^2$ or $(19n^2 - 41n + 24)/2$ or any of the infinitely many expressions of a similar type that could have been considered. The fact that the proof of a theorem consists in the application of certain simple rules of logic does not dispose of the creative element in mathematics, which lies in the choice of the possibilities to be examined. The question of the origin of the *hypothesis* (5) belongs to a domain in which no very general rules can be given; experiment, analogy, and constructive intuition play their part here. But once the correct hypothesis is formulated, the principle of mathematical induction is often sufficient to provide the proof. Inasmuch as such a proof does not give a clue to the act of discovery, it might more fittingly be called a *verification*.

*5. An Important Inequality

In a subsequent chapter we shall find use for the inequality

$$(1+p)^n \geq 1 + np, \tag{6}$$

which holds for every number $p > -1$ and positive integer n. (For the sake of generality we are anticipating here the use of negative and non-integral numbers by allowing p to be any number greater than -1. The proof for the general case is exactly the same as in the case where p is a positive integer.) Again we use mathematical induction.

a) *If* it is true that $(1+p)^r \geq 1 + rp$, then on multiplying both sides of this inequality by the positive number $1 + p$, we obtain

$$(1+p)^{r+1} \geq 1 + rp + p + rp^2.$$

Dropping the positive term rp^2 only strengthens this inequality, so that

$$(1+p)^{r+1} \geq 1 + (r+1)p,$$

Now subtract corresponding sides of this equation from the preceding equation, obtaining

$$G_n - qG_n = a - aq^{n+1},$$
$$(1 - q)G_n = a(1 - q^{n+1}),$$
$$G_n = a\frac{1 - q^{n+1}}{1 - q}.$$

4. The Sum of the First n Squares

A further interesting application of the principle of mathematical induction refers to the sum of the first n squares. By direct trial one finds that, at least for small values of n,

$$(4) \qquad 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

and one might *guess* that this remarkable formula is valid for *all integers* n. To *prove* this, we shall again use the principle of mathematical induction. We begin by observing that *if* the assertion A_n, which in this case is the equation (4), is true for the case $n = r$, so that

$$1^2 + 2^2 + 3^2 + \cdots + r^2 = \frac{r(r+1)(2r+1)}{6},$$

then on adding $(r+1)^2$ to both sides of this equation we obtain

$$1^2 + 2^2 + 3^2 + \cdots + r^2 + (r+1)^2 = \frac{r(r+1)(2r+1)}{6} + (r+1)^2$$
$$= \frac{r(r+1)(2r+1) + 6(r+1)^2}{6} = \frac{(r+1)[r(2r+1) + 6(r+1)]}{6}$$
$$= \frac{(r+1)(2r^2 + 7r + 6)}{6} = \frac{(r+1)(r+2)(2r+3)}{6},$$

which is precisely the assertion A_{r+1} in this case, since it is obtained by substituting $r + 1$ for n in (4). To complete the proof we need only remark that the assertion A_1, in this case the equation

$$1^2 = \frac{1(1+1)(2+1)}{6},$$

is obviously true. Hence the equation (4) is true for every n.

Formulas of a similar sort may be found for higher powers of the integers, $1^k + 2^k + 3^k + \cdots + n^k$, where k is any positive integer As an exercise, the reader may prove by mathematical induction that

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \left[\frac{n(n+1)}{2}\right]^2. \tag{5}$$

It should be remarked that although the principle of mathematical induction suffices to *prove* the formula (5) once this formula has been written down, the proof gives no indication of how this formula was arrived at in the first place; why precisely the expression $[n(n+1)/2]^2$ should be guessed as an expression for the sum of the first n cubes, rather than $[n(n+1)/3]^2$ or $(19n^2 - 41n + 24)/2$ or any of the infinitely many expressions of a similar type that could have been considered. The fact that the proof of a theorem consists in the application of certain simple rules of logic does not dispose of the creative element in mathematics, which lies in the choice of the possibilities to be examined. The question of the origin of the *hypothesis* (5) belongs to a domain in which no very general rules can be given; experiment, analogy, and constructive intuition play their part here. But once the correct hypothesis is formulated, the principle of mathematical induction is often sufficient to provide the proof. Inasmuch as such a proof does not give a clue to the act of discovery, it might more fittingly be called a *verification*.

*5. An Important Inequality

In a subsequent chapter we shall find use for the inequality

$$(1+p)^n \geq 1 + np, \tag{6}$$

which holds for every number $p > -1$ and positive integer n. (For the sake of generality we are anticipating here the use of negative and non-integral numbers by allowing p to be any number greater than -1. The proof for the general case is exactly the same as in the case where p is a positive integer.) Again we use mathematical induction.

a) *If* it is true that $(1+p)^r \geq 1 + rp$, then on multiplying both sides of this inequality by the positive number $1 + p$, we obtain

$$(1+p)^{r+1} \geq 1 + rp + p + rp^2.$$

Dropping the positive term rp^2 only strengthens this inequality, so that

$$(1+p)^{r+1} \geq 1 + (r+1)p,$$

which shows that the inequality (6) will also hold for the next integer, $r + 1$. b) It is obviously true that $(1 + p)^1 \geq 1 + p$. This completes the proof that (6) is true for every n. The restriction to numbers $p > -1$ is essential. If $p < -1$, then $1 + p$ is negative and the argument in a) breaks down, since if both members of an inequality are multiplied by a negative quantity, the sense of the inequality is reversed. (For example, if we multiply both sides of the inequality $3 > 2$ by -1 we obtain $-3 > -2$, which is false.)

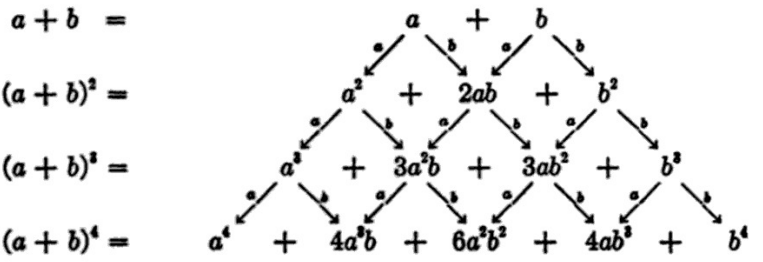
*6. The Binomial Theorem

Frequently it is important to have an explicit expression for the nth power of a binomial, $(a + b)^n$. We find by explicit calculation that for $n = 1$, $(a + b)^1 = a + b$,

for $n = 2$, $(a + b)^2 = (a + b)(a + b) = a(a + b) + b(a + b)$
$= a^2 + 2ab + b^2$,

for $n = 3$, $(a + b)^3 = (a + b)(a + b)^2 = a(a^2 + 2ab + b^2)$
$+ b(a^2 + 2ab + b^2) = a^3 + 3a^2b + 3ab^2 + b^3$,

and so on. What general law of formation lies behind the words "and so on"? Let us examine the process by which $(a + b)^2$ was computed. Since $(a + b)^2 = (a + b)(a + b)$, we obtained the expression for $(a + b)^2$ by multiplying each term in the expression $a + b$ by a, then by b, and adding. The same procedure was used to calculate $(a + b)^3 = (a + b)(a + b)^2$. We may continue in the same way to calculate $(a + b)^4$, $(a + b)^5$, and so on indefinitely. The expression for $(a + b)^n$ will be obtained by multiplying each term of the previously obtained expression for $(a + b)^{n-1}$ by a, then by b, and adding. This leads to the following diagram:



which gives at once the general rule for forming the coefficients in the expansion of $(a + b)^n$. We construct a triangular array of numbers,

starting with the coefficients 1, 1 of $a + b$, and such that each number of the triangle is the sum of the two numbers on each side of it in the preceding row. This array is known as *Pascal's Triangle.*

$$
\begin{array}{ccccccccccccc}
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1
\end{array}
$$

.....

The nth row of this array gives the coefficients in the expansion of $(a + b)^n$ *in descending powers of* a *and ascending powers of* b; thus

$$(a + b)^7 = a^7 + 7a^6b + 21a^5b^2 + 35a^4b^3 + 35a^3b^4 + 21a^2b^5 + 7ab^6 + b^7.$$

Using a concise subscript and superscript notation we may denote the numbers in the nth row of Pascal's Triangle by

$$C_0^n = 1, C_1^n, C_2^n, C_3^n, \cdots, C_{n-1}^n, C_n^n = 1.$$

Then the general formula for $(a + b)^n$ may be written

$$(7) \quad (a + b)^n = a^n + C_1^n a^{n-1}b + C_2^n a^{n-2}b^2 + \cdots + C_{n-1}^n ab^{n-1} + b^n.$$

According to the law of formation of Pascal's Triangle, we have

$$(8) \qquad C_i^n = C_{i-1}^{n-1} + C_i^{n-1}.$$

As an exercise, the experienced reader may use this relation, together with the fact that $C_0^1 = C_1^1 = 1$, to show by mathematical induction that

$$(9) \qquad C_i^n = \frac{n(n-1)(n-2)\cdots(n-i+1)}{1\cdot 2\cdot 3\cdots i} = \frac{n!}{i!\,(n-i)!}.$$

(For any positive integer n, the symbol $n!$ (read, "n factorial") denotes the product of the first n integers: $n! = 1\cdot 2\cdot 3\cdots n$. It is convenient also to define $0! = 1$, so that 9) is valid for $i = 0$ and $i = n$.) This explicit formula for the coefficients in the binomial expansion is sometimes called the *binomial theorem.* (See also p. 475.)

Exercises: Prove by mathematical induction:

1) $\dfrac{1}{1\cdot 2} + \dfrac{1}{2\cdot 3} + \cdots + \dfrac{1}{n(n+1)} = \dfrac{n}{n+1}.$

2) $\dfrac{1}{2} + \dfrac{2}{2^2} + \dfrac{3}{2^3} + \cdots + \dfrac{n}{2^n} = 2 - \dfrac{n+2}{2^n}.$

*3) $1 + 2q + 3q^2 + \cdots + nq^{n-1} = \dfrac{1 - (n+1)q^n + nq^{n+1}}{(1-q)^2}$.

*4) $(1+q)(1+q^2)(1+q^4)\cdots(1+q^{2^n}) = \dfrac{1 - q^{2^{n+1}}}{1-q}$.

Find the sum of the following geometrical progressions:

5) $\dfrac{1}{1+x^2} + \dfrac{1}{(1+x^2)^2} + \cdots + \dfrac{1}{(1+x^2)^n}$.

6) $1 + \dfrac{x}{1+x^2} + \dfrac{x^2}{(1+x^2)^2} + \cdots + \dfrac{x^n}{(1+x^2)^n}$.

7) $\dfrac{x^2-y^2}{x^2+y^2} + \left(\dfrac{x^2-y^2}{x^2+y^2}\right)^2 + \cdots + \left(\dfrac{x^2-y^2}{x^2+y^2}\right)^n$.

Using formulas (4) and (5) prove:

*8) $1^2 + 3^2 + \cdots + (2n+1)^2 = \dfrac{(n+1)(2n+1)(2n+3)}{3}$.

*9) $1^3 + 3^3 + \cdots + (2n+1)^3 = (n+1)^2(2n^2+4n+1)$.

10) Prove the same results directly by mathematical induction.

*7. Further Remarks on Mathematical Induction

The principle of mathematical induction may be generalized slightly to read: "If a sequence of statements A_s, A_{s+1}, A_{s+2}, $\cdots$ is given, where s is some positive integer, and if

a) For every value of $r \geq s$, the truth of A_{r+1} will follow from the truth of A_r, and

b) A_s is known to be true,

then all the statements A_s, A_{s+1}, A_{s+2} $\cdots$ are true; that is to say, A_n is true for all $n \geq s$." Precisely the same reasoning used to establish the truth of the ordinary principle of mathematical induction applies here, with the sequence 1, 2, 3, $\cdots$ replaced by the similar sequence s, $s+1$, $s+2$, $s+3$ $\cdots$. By using the principle in this form we can strengthen somewhat the inequality on page 15 by eliminating the possibility of the "=" sign. We state: *For every* p $\neq 0$ *and* > -1 *and every integer* n ≥ 2,

$$(1+p)^n > 1 + np. \tag{10}$$

The proof will be left to the reader.

Closely related to the principle of mathematical induction is the "principle of the smallest integer" which states that *every non-empty set* C *of positive integers has a smallest member.* A set is empty if it has no members, e.g., the set of straight circles or the set of integers n such that $n > n$. For obvious reasons we exclude such sets in the statement of the principle. The set C may be finite, like the set 1, 2, 3, 4, 5, or infinite, like the set of all even numbers 2, 4, 6, 8, 10, $\cdots$. Any non-empty set C must contain at least one integer, say n, and the smallest of the integers 1, 2, 3, $\cdots$, n that belongs to C will be the smallest integer in C.

The only way to realize the significance of this principle is to observe that it

does *not* apply to every set C of numbers that are not integers; for example, the set of positive fractions $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots$ does not contain a smallest member.

From the point of view of logic it is interesting to observe that the principle of the smallest integer may be used to *prove* the principle of mathematical induction as a theorem. To this end, let us consider any sequence of statements $A_1, A_2, A_3, \cdots$ such that

a) For any positive integer r the truth of A_{r+1} will follow from that of A_r.

b) A_1 is known to be true.

We shall show the hypothesis that any one of the A's is false to be untenable. For if even one of the A's were false, the set C of *all* positive integers n for which A_n is false would be non-empty. By the principle of the smallest integer, C would contain a smallest integer, p, which must be > 1 because of b). Hence A_p would be false, but A_{p-1} true. This contradicts a)..

Once more we emphasize that the principle of mathematical induction is quite distinct from empirical induction in the natural sciences. The confirmation of a general law in any finite number of cases, no matter how large, cannot provide a proof for the law in the rigorous mathematical sense of the word, even if no exception is known at the time. Such a law would remain only a very reasonable *hypothesis*, subject to modification by the results of future experience. In mathematics, a law or a theorem is proved only if it can be shown to be a necessary logical consequence of certain assumptions which are accepted as valid. There are many examples of mathematical statements which have been verified in every particular case considered thus far, but which have not yet been proved to hold in general (for an example see p. 30). One may *suspect* that a theorem is true in all generality by observing its truth in a number of examples; one may then attempt to *prove* it by mathematical induction. If the attempt succeeds the theorem is proved to be true; if the attempt fails, the theorem may be true or false and may some day be proved or disproved by other methods.

In using the principle of mathematical induction one must always be sure that the conditions a) and b) are really satisfied. Neglect of this precaution may lead to absurdities like the following, in which the reader is invited to discover the fallacy. We shall "prove" that *any two positive integers are equal*; for example, that $5 = 10$.

First a definition: If a and b are two unequal positive integers, we define $\max(a, b)$ to be a or b, whichever is greater; if $a = b$ we set $\max(a, b) = a = b$. Thus $\max(3, 5) = \max(5, 3) = 5$, while $\max(4, 4) = 4$. Now let A_n be the statement, "If a and b are any two positive integers such that $\max(a, b) = n$, then $a = b$."

a) Suppose A_r to be true. Let a and b be any two positive integers such that $\max(a, b) = r + 1$. Consider the two integers

$$\alpha = a - 1$$

$$\beta = b - 1;$$

then $\max(\alpha, \beta) = r$. Hence $\alpha = \beta$, for we are assuming A_r to be true. It follows that $a = b$; hence A_{r+1} is true.

b) A_1 is obviously true, for if $\max(a, b) = 1$, then since a and b are by hypothesis positive integers they must both be equal to 1. Therefore, by mathematical induction, A_n is true for every n.

Now if a and b are any two positive integers whatsoever, denote $\max(a, b)$ by r. Since A_n has been shown to be true for every n, in particular A_r is true. Hence $a = b$.

SUPPLEMENT TO CHAPTER I

THE THEORY OF NUMBERS

INTRODUCTION

The integers have gradually lost their association with superstition and mysticism, but their interest for mathematicians has never waned. Euclid (circa 300 B.C.), whose fame rests on the portion of his *Elements* that forms the foundation of geometry studied in high school, seems to have made original contributions to number theory, while his geometry was largely a compilation of previous results. Diophantus of Alexandria (circa 275 A.D.), an early algebraist, left his mark on the theory of numbers. Pierre de Fermat (1601–1665), a jurist of Toulouse, and one of the greatest mathematicians of his time, initiated the modern work in this field. Euler (1707–1783), the most prolific of mathematicians, included much number-theoretical work in his researches. Names prominent in the annals of mathematics—Legendre, Dirichlet, Riemann—can be added to the list. Gauss (1777–1855), the foremost mathematician of modern times, who devoted himself to many different branches of mathematics, is said to have expressed his opinion of number theory in the remark, "Mathematics is the queen of the sciences and the theory of numbers is the queen of mathematics."

§1. THE PRIME NUMBERS

1. Fundamental Facts

Most statements in number theory, as in mathematics as a whole, are concerned not with a single object—the number 5 or the number 32—but with a whole class of objects that have some common property, such as the class of all even integers,

$$2, 4, 6, 8, \cdots,$$

or the class of all integers divisible by 3,

$$3, 6, 9, 12, \cdots,$$

or the class of all squares of integers,

$$1, 4, 9, 16, \cdots,$$

and so on.

either $p_1 < q_1$ or $q_1 < p_1$. Suppose $p_1 < q_1$. (If $q_1 < p_1$ we simply interchange the letters p and q in what follows.) We form the integer

$$m' = m - (p_1q_2q_3 \cdots q_s). \tag{2}$$

By substituting for m the two expressions of equation (1) we may write the integer m' in either of the two forms

$$m' = (p_1p_2 \cdots p_r) - (p_1q_2 \cdots q_s) = p_1(p_2p_3 \cdots p_r - q_2q_3 \cdots q_s) \tag{3}$$

$$m' = (q_1q_2 \cdots q_s) - (p_1q_2 \cdots q_s) = (q_1 - p_1)(q_2q_3 \cdots q_s) \tag{4}$$

Since $p_1 < q_1$, it follows from (4) that m' is a positive integer, while from (2) it follows that m' is smaller than m. Hence the prime decomposition of m' must be *unique*, aside from the order of the factors. But from (3) it appears that the prime p_1 is a factor of m', hence from (4) p_1 must appear as a factor of either $(q_1 - p_1)$ or $(q_2q_3 \cdots q_s)$. (This follows from the assumed uniqueness of the prime decomposition of m'; see the reasoning in the next paragraph.) The latter is impossible, since all the q's are larger than p_1. Hence p_1 must be a factor of $q_1 - p_1$, so that for some integer h,

$$q_1 - p_1 = p_1 \cdot h \qquad \text{or} \qquad q_1 = p_1(h + 1).$$

But this shows that p_1 is a factor of q_1, contrary to the fact that q_1 is a prime. This contradiction shows our initial assumption to be untenable and hence completes the proof of the fundamental theorem of arithmetic.

An important corollary of the fundamental theorem is the following: *If a prime* p *is a factor of the product* ab, *then* p *must be a factor of either* a *or* b. For if p were a factor of neither a nor b, then the product of the prime decompositions of a and b would yield a prime decomposition of the integer *ab not containing p*. On the other hand, since p is assumed to be a factor of ab, there exists an integer t such that

$$ab = pt.$$

Hence the product of p by a prime decomposition of t would yield a prime decomposition of the integer *ab containing p*, contrary to the fact that the prime decomposition of ab is unique.

Examples: If one has verified the fact that 13 is a factor of 2652, and the fact that $2652 = 6 \cdot 442$, one may conclude that 13 is a factor of 442. On the other hand, 6 is a factor of 240, and $240 = 15 \cdot 16$, but 6 is not a factor of either 15 or 16. This shows that the assumption that p is *prime* is an essential one.

Exercise: In order to find all the divisors of any number a we need only decompose a into a product

$$a = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_r^{\alpha_r},$$

where the p's are distinct primes, each raised to a certain power. *All* the divisors of a are the numbers

$$b = p_1^{\beta_1} \cdot p_2^{\beta_2} \cdots p^{\beta_r},$$

where the β's are any integers satisfying the inequalities

$$0 \leq \beta_1 \leq \alpha_1, \; 0 \leq \beta_2 \leq \alpha_2, \; \cdots, \; 0 \leq \beta_r \leq \alpha_r.$$

Prove this statement. As a consequence, show that the number of different divisors of a (including the divisors a and 1) is given by the product

$$(\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_r + 1).$$

For example,

$$144 = 2^4 \cdot 3^2$$

has $5 \cdot 3$ divisors. They are 1, 2, 4, 8, 16, 3, 6, 12, 24, 48, 9, 18, 36, 72, 144.

2. The Distribution of the Primes

A list of all the primes up to any given integer N may be constructed by writing down in order all the integers less than N, striking out all those which are multiples of 2, then all those remaining which are multiples of 3, and so on until all composite numbers have been eliminated. This process, known as the "sieve of Eratosthenes," will catch in its meshes the primes up to N. Complete tables of primes up to about 10,000,000 have gradually been computed by refinements of this method, and they provide us with a tremendous mass of empirical data concerning the distribution and properties of the primes. On the basis of these tables we can make many highly plausible conjectures (as though number theory were an experimental science) which are often extremely difficult to prove.

a. Formulas Producing Primes

Attempts have been made to find simple arithmetical formulas that yield only primes, even though they may not give all of them. Fermat made the famous conjecture (but not the definite assertion) that all numbers of the form

$$F(n) = 2^{2^n} + 1$$

are primes. Indeed, for $n = 1, 2, 3, 4$ we obtain

$$F(1) = 2^2 + 1 = 5,$$
$$F(2) = 2^{2^2} + 1 = 2^4 + 1 = 17,$$
$$F(3) = 2^{2^3} + 1 = 2^8 + 1 = 257,$$
$$F(4) = 2^{2^4} + 1 = 2^{16} + 1 = 65{,}537,$$

all primes. But in 1732 Euler discovered the factorization $2^{2^5} + 1 = 641 \cdot 6{,}700{,}417$; hence $F(5)$ is not a prime. Later, more of these "Fermat numbers" were found to be composite, deeper number-theoretical methods being required in each case because of the insurmountable difficulty of direct trial. To date it has not even been proved that any of the numbers $F(n)$ is a prime for $n > 4$.

Another remarkable and simple expression which produces many primes is

$$f(n) = n^2 - n + 41.$$

For $n = 1, 2, 3, \cdots, 40$, $f(n)$ is a prime; but for $n = 41$, we have $f(n) = 41^2$, which is no longer a prime.

The expression

$$n^2 - 79n + 1601$$

yields primes for all n up to 79, but fails when $n = 80$. On the whole, it has been a futile task to seek expressions of a simple type which produce only primes. Even less promising is the attempt to find an algebraic formula which shall yield *all* the primes.

b. Primes in Arithmetical Progressions

While it was simple to prove that there are infinitely many primes in the sequence of all integers, 1, 2, 3, 4, $\cdots$, the step to sequences such as 1, 4, 7, 10, 13, $\cdots$ or 3, 7, 11, 15, 19, $\cdots$ or, more generally, to any arithmetical progression, $a, a + d, a + 2d, \cdots a + nd, \cdots$, where a and d have no common factor, was much more difficult. All observations pointed to the fact that *in each such progression there are infinitely many primes*, just as in the simplest one, 1, 2, 3, $\cdots$. It required an enormous effort to prove this general theorem. Lejeune Dirichlet (1805–1859), one of the leading mathematicians of the nineteenth century, obtained full success by applying the most advanced tools of mathematical analysis then known. His original papers on the subject rank even now among the outstanding achievements in mathematics, and after a hundred years the proof has not yet been simplified enough to be within the reach of students who are not well trained in the technique of the calculus and of function theory.

Although we cannot attempt to prove Dirichlet's general theorem, it is easy to generalize Euclid's proof of the infinitude of primes to cover some *special* arithmetical progressions such as $4n + 3$ and $6n + 5$. To treat the first of these, we observe that any prime greater than 2 is odd (since otherwise it would be divisible by 2) and hence is of the form $4n + 1$ or $4n + 3$, for some integer n. Furthermore, the product of two numbers of the form $4n + 1$ is again of that form, since

$$(4a + 1)(4b + 1) = 16ab + 4a + 4b + 1 = 4(4ab + a + b) + 1.$$

Now suppose there were but a finite number of primes, $p_1, p_1, \cdots p_n$, of the form $4n + 3$, and consider the number

$$N = 4(p_1p_2 \cdots p_n) - 1 = 4(p_1 \cdots p_n - 1) + 3.$$

Either N is itself a prime, or it may be decomposed into a product of primes, none of which can be $p_1, \cdots, p_n$, since these divide N with a remainder -1. Furthermore, all the prime factors of N cannot be of the form $4n + 1$, for N is not of that form and, as we have seen, the product of numbers of the form $4n + 1$ is again of that form. Hence at least one prime factor must be of the form $4n + 3$, which is impossible, since we saw that none of the p's, which we supposed to be *all* the primes of the form $4n + 3$, can be a factor of N. Therefore the assumption that the number of primes of the form $4n + 3$ is finite has led to a contradiction, and hence the number of such primes must be infinite.

Exercise: Prove the corresponding theorem for the progression $6n + 5$.

c. *The Prime Number Theorem*

In the search for a law governing the distribution of the primes, the decisive step was taken when mathematicians gave up futile attempts to find a simple mathematical formula yielding all the primes or giving the exact number of primes contained among the first n integers, and sought instead for information concerning the *average* distribution of the primes among the integers.

For any integer n let us denote by A_n the number of primes among the integers $1, 2, 3, \cdots, n$. If we underline the primes in the sequence consisting of the first few integers: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 $\cdots$ we can compute the first few values of A_n :

$A_1 = 0, A_2 = 1, A_3 = A_4 = 2, A_5 = A_6 = 3, A_7 = A_8 = A_9 = A_{10} = 4,$
$A_{11} = A_{12} = 5, A_{13} = A_{14} = A_{15} = A_{16} = 6, A_{17} = A_{18} = 7, A_{19} = 8,$ etc.

If we now take any sequence of values for n which increases without limit, say

$$n = 10, 10^2, 10^3, 10^4, \cdots,$$

then the corresponding values of A_n,

$$A_{10}, A_{10^2}, A_{10^3}, A_{10^4}, \cdots,$$

will also increase without limit (although more slowly). For we know that there are infinitely many primes, so the values of A_n will sooner or later exceed any finite number. The "density" of the primes among the first n integers is given by the ratio A_n/n, and from a table of primes the values of A_n/n may be computed empirically for fairly large values of n.

n	A_n/n
10^3	0.168
10^6	0.078498
10^9	0.050847478
...	.....

The last entry in this table may be regarded as giving the probability that an integer picked at random from among the first 10^9 integers will be a prime, since there are 10^9 possible choices, of which A_{10^9} are primes.

The distribution of the individual primes among the integers is extremely irregular. But this irregularity "in the small" disappears if we fix our attention on the average distribution of the primes as given by the ratio A_n/n. The simple law that governs the behavior of this ratio is one of the most remarkable discoveries in the whole of mathematics. In order to state the *prime number theorem* we must define the "natural logarithm" of an integer n. To do this we take two perpendicular axes in a plane, and consider the locus of all points in the plane the product of whose distances x and y from these axes is equal to one. In terms of the coördinates x and y this locus, an equilateral hyperbola, is defined by the equation $xy = 1$. We now define log n to be the *area* in Figure 5 bounded by the hyperbola, the x-axis, and the two vertical lines $x = 1$ and $x = n$. (A more detailed discussion of the logarithm will be found in Chapter VIII.) From an empirical study of prime number tables Gauss observed that the ratio A_n/n is approximately equal to $1/\log n$, and that this approximation appears to improve

as n increases. The goodness of the approximation is given by the ratio $\dfrac{A_n/n}{1/\log n}$, whose values for $n = 1000, 1{,}000{,}000, 1{,}000{,}000{,}000$ are shown in the following table.

n	A_n/n	$1/\log n$	$\dfrac{A_n/n}{1/\log n}$
10^3	0.168	0.145	1.159
10^6	0.078498	0.072382	1.084
10^9	0.050847478	0.048254942	1.053
...	.....	.....	.....

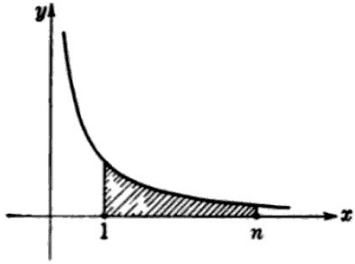


Fig. 5. The area of the shaded region under the hyperbola defines log n.

On the basis of such empirical evidence Gauss made the conjecture that the ratio A_n/n is "asymptotically equal" to $1/\log n$. By this is meant that if we take a sequence of larger and larger values of n, say n equal to

$$10, 10^2, 10^3, 10^4, \ldots$$

as before, then the ratio of A_n/n to $1/\log n$,

$$\frac{A_n/n}{1/\log n},$$

calculated for these successive values of n, will become more and more nearly equal to 1, and that the difference of this ratio from 1 can be made as small as we please by confining ourselves to sufficiently large values of n. This assertion is symbolically expressed by the sign $\sim$:

$$\frac{A_n}{n} \sim \frac{1}{\log n} \quad \text{means} \quad \frac{A_n/n}{1/\log n} \text{ tends to 1 as } n \text{ increases.}$$

That $\sim$ cannot be replaced by the ordinary sign $=$ of equality is clear from the fact that while A_n is always an integer, $n/\log n$ is not.

That the average behavior of the prime number distribution can be described by the logarithmic function is a very remarkable discovery, for it is surprising that two mathematical concepts which seem so unrelated should be in fact so intimately connected.

Although the statement of Gauss's conjecture is simple to understand, a rigorous mathematical proof was far beyond the powers of mathematical science in Gauss's time. To prove this theorem, concerned only with the most elementary concepts, it is necessary to employ the most powerful methods of modern mathematics. It took almost a hundred years before analysis was developed to the point where Hadamard (1896) in Paris and de la Vallée Poussin (1896) in Louvain could give a complete proof of the prime number theorem. Simplifications and important modifications were given by v. Mangoldt and Landau. Long before Hadamard, decisive pioneering work had been done by Riemann (1826–1866) in a famous paper where the strategic lines for the attack were drawn. Recently, the American mathematician Norbert Wiener was able to modify the proof so as to avoid the use of complex numbers at an important step of the reasoning. But the proof of the prime number theorem is still no easy matter even for an advanced student. We shall return to this subject on page 482 et seq.

d. Two Unsolved Problems Concerning Prime Numbers

While the problem of the average distribution of primes has been satisfactorily solved, there are many other conjectures which are supported by all the empirical evidence but which have not yet been proved to be true.

One of these is the famous *Goldbach conjecture.* Goldbach (1690–1764) has no significance in the history of mathematics except for this problem, which he proposed in 1742 in a letter to Euler. He observed that for every case he tried, any even number (except 2, which is itself a prime) could be represented as the sum of two primes. For example:

$4 = 2 + 2, 6 = 3 + 3, 8 = 5 + 3, 10 = 5 + 5, 12 = 5 + 7, 14 = 7 + 7, 16 = 13 + 3, 18 = 11 + 7, 20 = 13 + 7, \cdots, 48 = 29 + 19, \cdots, 100 = 97 + 3,$ etc.

Goldbach asked if Euler could prove this to be true for *all* even numbers, or if he could find an example disproving it. Euler never provided an answer, nor has one been given since. The empirical evidence in favor of the statement that every even number can be so represented is thoroughly convincing, as anyone can verify by trying a number of examples. The source of the difficulty is that primes are defined in terms of *multiplication*, while the problem involves *addition*. Generally

speaking, it is difficult to establish connections between the multiplicative and the additive properties of integers.

Until recently, a proof of Goldbach's conjecture seemed completely inaccessible. Today a solution no longer seems out of reach. An important success, very unexpected and startling to all experts, was achieved in 1931 by a then unknown young Russian mathematician, Schnirelmann (1905–1938), who proved that *every positive integer can be represented as the sum of not more than 300,000 primes.* Though this result seems ludicrous in comparison with the original goal of proving Goldbach's conjecture, nevertheless it was a first step in that direction. The proof is a direct, constructive one, although it does not provide any practical method for finding the prime decomposition of an arbitrary integer. More recently, the Russian mathematician Vinogradoff, using methods due to Hardy, Littlewood and their great Indian collaborator Ramanujan, has succeeded in reducing the number from 300,000 to 4. This is much nearer to a solution of Goldbach's problem. But there is a striking difference between Schnirelmann's result and Vinogradoff's; more significant, perhaps, than the difference between 300,000 and 4. Vinogradoff's theorem was proved only fór all "sufficiently large" integers; more precisely, Vinogradoff proved that there *exists* an integer N such that any integer $n > N$ can be represented as the sum of at most 4 primes. Vinogradoff's proof does not permit us to appraise N; in contrast to Schnirelmann's theorem it is essentially indirect and non-constructive. What Vinogradoff really proved is that the assumption that infinitely many integers cannot be decomposed into at most 4 prime summands leads to an absurdity. Here we have a good example of the profound difference between the two types of proof, direct and indirect. (See the general discussion on p. 86.)

The following even more striking problem than Goldbach's has come nowhere near a solution. It has been observed that primes frequently occur in pairs of the form p and $p + 2$. Such are 3 and 5, 11 and 13, 29 and 31, etc. The statement that there are infinitely many such pairs is believed to be correct, but as yet not the slightest definite step has been taken towards a proof.

§2. CONGRUENCES

1. General Concepts

Whenever the question of the divisibility of integers by a fixed integer d occurs, the concept and the notation of "congruence" (due to Gauss) serves to clarify and simplify the reasoning.

To introduce this concept let us examine the remainders left when integers are divided by the number 5. We have

$0 = 0 \cdot 5 + 0$	$7 = 1 \cdot 5 + 2$	$-1 = -1 \cdot 5 + 4$
$1 = 0 \cdot 5 + 1$	$8 = 1 \cdot 5 + 3$	$-2 = -1 \cdot 5 + 3$
$2 = 0 \cdot 5 + 2$	$9 = 1 \cdot 5 + 4$	$-3 = -1 \cdot 5 + 2$
$3 = 0 \cdot 5 + 3$	$10 = 2 \cdot 5 + 0$	$-4 = -1 \cdot 5 + 1$
$4 = 0 \cdot 5 + 4$	$11 = 2 \cdot 5 + 1$	$-5 = -1 \cdot 5 + 0$
$5 = 1 \cdot 5 + 0$	$12 = 2 \cdot 5 + 2$	$-6 = -2 \cdot 5 + 4$
$6 = 1 \cdot 5 + 1$	etc.	etc.

We observe that the remainder left when any integer is divided by 5 is one of the five integers 0, 1, 2, 3, 4. We say that two integers a and b are "congruent modulo 5" if they leave the *same remainder* on division by 5. Thus 2, 7, 12, 17, 22, $\cdots$, -3, -8, -13, -18, $\cdots$ are all congruent modulo 5, since they leave the remainder 2. In general, we say that two integers a and b are *congruent modulo* d, where d is a fixed integer, if a and b leave the same remainder on division by d, i.e., if there is an integer n such that $a - b = nd$. For example, 27 and 15 are congruent modulo 4, since

$$27 = 6 \cdot 4 + 3, \qquad 15 = 3 \cdot 4 + 3.$$

The concept of congruence is so useful that it is desirable to have a brief notation for it. We write

$$a \equiv b \pmod{d}$$

to express the fact that a and b are congruent modulo d. If there is no doubt concerning the modulus, the "mod d" of the formula may be omitted. (If a is not congruent to b modulo d, we shall write $a \not\equiv b \pmod{d}$.)

Congruences occur frequently in daily life. For example, the hands on a clock indicate the hour modulo 12, and the mileage indicator on a car gives the total miles traveled modulo 100,000.

Before proceeding with the detailed discussion of congruences the reader should observe that the following statements are all equivalent:

1. a is congruent to b modulo d.
2. $a = b + nd$ for some integer n.
3. d divides $a - b$.

The usefulness of Gauss's congruence notation lies in the fact that congruence with respect to a fixed modulus has many of the formal

properties of ordinary equality. The most important formal properties of the relation $a = b$ are the following:

1) Always $a = a$.
2) If $a = b$, then $b = a$.
3) If $a = b$ and $b = c$, then $a = c$.

Moreover, if $a = a'$ and $b = b'$, then

4) $a + b = a' + b'$.
5) $a - b = a' - b'$.
6) $ab = a'b'$.

These properties remain true when the relation $a = b$ is replaced by the congruence relation $a \equiv b \pmod{d}$. Thus

1′) Always $a \equiv a \pmod{d}$.
2′) If $a \equiv b \pmod{d}$ then $b \equiv a \pmod{d}$.
3′) If $a \equiv b \pmod{d}$ and $b \equiv c \pmod{d}$, then $a \equiv c \pmod{d}$.

The trivial verification of these facts is left to the reader.

Moreover, if $a \equiv a' \pmod{d}$ and $b \equiv b' \pmod{d}$, then

4′) $a + b \equiv a' + b' \pmod{d}$.
5′) $a - b \equiv a' - b' \pmod{d}$.
6′) $ab \equiv a'b' \pmod{d}$.

Thus *congruences with respect to the same modulus may be added, subtracted, and multiplied.* To prove these three statements we need only observe that if

$$a = a' + rd, \qquad b = b' + sd,$$

then

$$a + b = a' + b' + (r + s)d,$$
$$a - b = a' - b' + (r - s)d,$$
$$ab = a'b' + (a's + b'r + rsd)d,$$

from which the desired conclusions follow.

The concept of congruence has an illuminating geometrical interpretation. Usually, if we wish to represent the integers geometrically, we choose a segment of unit length and extend it by multiples of its own length in both directions. In this way we can find a point on the line corresponding to each integer, as in Figure 6. But when we are dealing with the integers modulo d, any two congruent numbers are considered the same as far as their behavior on division by a is concerned,

since they leave the same remainder. In order to show this geometrically, we use a circle divided into d equal parts. Any integer when divided by d leaves as remainder one of the d numbers $0, 1, \cdots, d-1$, which are placed at equal intervals on the circumference of the circle. Every integer is congruent modulo d to one of these numbers, and hence is represented geometrically by one of these points; two numbers are congruent if they are represented by the same point. Figure 7 is drawn for the case $d = 6$. The face of a clock is another illustration from daily life.

Fig. 6. Geometrical representation of the integers.

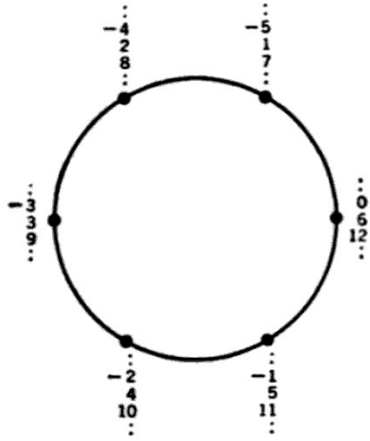


Fig. 7. Geometrical representation of the integers modulo 6.

As an example of the use of the multiplicative property 6′) of congruences we may determine the remainders left when successive powers of 10 are divided by a given number. For example,

$$10 \equiv -1 \pmod{11},$$

since $10 = -1 + 11$. Successively multiplying this congruence by itself, we obtain

$$10^2 \equiv (-1)(-1) = 1 \pmod{11},$$
$$10^3 \equiv -1 \quad \text{“} \quad ,$$
$$10^4 \equiv 1 \quad \text{“} \quad , \text{ etc.}$$

From this we can show that any integer

$$z = a_0 + a_1 \cdot 10 + a_2 \cdot 10^2 + \cdots + a_n \cdot 10^n,$$

expressed in the decimal system, leaves the same remainder on division by 11 as does the sum of its digits, taken with alternating signs,

$$t = a_0 - a_1 + a_2 - a_3 + \cdots .$$

For we may write

$$z - t = a_1 \cdot 11 + a_2(10^2 - 1) + a_3(10^3 + 1) + a_4(10^4 - 1) + \cdots .$$

Since all the numbers $11, 10^2 - 1, 10^3 + 1, \cdots$ are congruent to 0 modulo 11, $z - t$ is also, and therefore z leaves the same remainder on division by 11 as does t. It follows in particular that a number is divisible by 11 (i.e. leaves the remainder 0) if and only if the alternating sum of its digits is divisible by 11. For example, since $3 - 1 + 6 - 2 + 8 - 1 + 9 = 22$, the number $z = 3162819$ is divisible by 11. To find a rule for divisibility by 3 or 9 is even simpler, since $10 \equiv 1 \pmod{3 \text{ or } 9}$, and therefore $10^n \equiv 1 \pmod{3 \text{ or } 9}$ for any n. It follows that a number z is divisible by 3 or 9 if and only if the sum of its digits

$$s = a_0 + a_1 + a_2 + \cdots + a_n$$

is likewise divisible by 3 or 9, respectively.

For congruences modulo 7 we have

$$10 \equiv 3, \quad 10^2 \equiv 2, \quad 10^3 \equiv -1, \quad 10^4 \equiv -3, \quad 10^5 \equiv -2, \quad 10^6 \equiv 1.$$

The successive remainders then repeat. Thus z is divisible by 7 if and only if the expression

$$r = a_0 + 3a_1 + 2a_2 - a_3 - 3a_4 - 2a_5 + a_6 + 3a_7 + \cdots$$

is divisible by 7.

Exercise: Find a similar rule for divisibility by 13.

In adding or multiplying congruences with respect to a fixed modulus, say $d = 5$, we may keep the numbers involved from getting too large by always replacing any number a by the number from the set

$$0, \quad 1, \quad 2, \quad 3, \quad 4$$

to which it is congruent. Thus, in order to calculate sums and products of integers modulo 5, we need only use the following addition and multiplication tables.

$a + b$

	$b \equiv 0$	1	2	3	4
$a \equiv 0$	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$a \cdot b$

	$b \equiv 0$	1	2	3	4
$a \equiv 0$	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

From the second of these tables it appears that a product ab is congruent to 0 (mod 5) only if a or b is $\equiv 0 \pmod 5$. This suggests the general law

7) $ab \equiv 0 \pmod d$ only if either $a \equiv 0$ or $b \equiv 0 \pmod d$,

which is an extension of the ordinary law for integers which states that $ab = 0$ only if $a = 0$ or $b = 0$. *The law* 7) *holds only when the modulus* d *is a prime.* For the congruence

$$ab \equiv 0 \pmod d$$

means that d divides ab, and we have seen that a prime d divides a product ab only if it divides a or b; that is, only if

$$a \equiv 0 \pmod d \quad \text{or} \quad b \equiv 0 \pmod d.$$

If d is not a prime the law need not hold; for we can write $d = r \cdot s$, where r and s are less than d, so that

$$r \not\equiv 0 \pmod d, \qquad s \not\equiv 0 \pmod d,$$

but

$$rs = d \equiv 0 \pmod d.$$

For example, $2 \not\equiv 0 \pmod 6$ and $3 \not\equiv 0 \pmod 6$, but $2 \cdot 3 = 6 \equiv 0 \pmod 6$.

Exercise: Show that the following *law of cancellation* holds for congruences with respect to a prime modulus:

If $ab \equiv ac$ and $a \not\equiv 0$, then $b \equiv c$.

Exercises: 1) To what number between 0 and 6 inclusive is the product $11 \cdot 18 \cdot 2322 \cdot 13 \cdot 19$ congruent modulo 7?

2) To what number between 0 and 12 inclusive is $3 \cdot 7 \cdot 11 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 113$ congruent modulo 13?

3) To what number between 0 and 4 inclusive is the sum $1 + 2 + 2^2 + \cdots + 2^{19}$ congruent modulo 5?

2. Fermat's Theorem

In the seventeenth century, Fermat, the founder of modern number theory, discovered a most important theorem: *If* p *is any prime which does not divide the integer* a, *then*

$$a^{p-1} \equiv 1 \pmod{p}.$$

This means that the $(p-1)$st power of a leaves the remainder 1 upon division by p.

Some of our previous calculations confirm this theorem; for example, we found that $10^6 \equiv 1 \pmod 7$, $10^2 \equiv 1 \pmod 3$, and $10^{10} \equiv 1 \pmod{11}$. Likewise we may show that $2^{12} \equiv 1 \pmod{13}$ and $5^{10} \equiv 1 \pmod{11}$. To check the latter congruences we need not actually calculate such high powers, since we may take advantage of the multiplicative property of congruences:

$2^4 = 16 \equiv 3 \pmod{13}$, $\quad 5^2 \equiv 3 \pmod{11}$,

$2^8 \equiv 9 \equiv -4$ " , $\quad 5^4 \equiv 9 \equiv -2$ " ,

$2^{12} \equiv -4 \cdot 3 = -12 \equiv 1$ " . $\quad 5^8 \equiv 4$ " ,

$5^{10} \equiv 3 \cdot 4 = 12 \equiv 1$ " .

To prove Fermat's theorem, we consider the multiples of a

$$m_1 = a, \quad m_2 = 2a, \quad m_3 = 3a, \cdots, m_{p-1} = (p-1)a.$$

No two of these integers can be congruent modulo p, for then p would be a factor of $m_r - m_s = (r-s)a$ for some pair of integers r, s with $1 \leq r < s \leq (p-1)$. But the law 7) shows that this cannot occur; for since $s - r$ is less than p, p is not a factor of $s - r$, while by assumption p is not a factor of a. Likewise, none of these numbers can be congruent to 0. Therefore the numbers $m_1, m_2, \cdots, m_{p-1}$ must be respectively congruent to the numbers $1, 2, 3, \cdots, p-1$, in some arrangement. It follows that

$$m_1 m_2 \cdots m_{p-1} = 1 \cdot 2 \cdot 3 \cdots (p-1)a^{p-1} \equiv 1 \cdot 2 \cdot 3 \cdots (p-1) \pmod{p},$$

or, if for brevity we write K for $1 \cdot 2 \cdot 3 \cdots (p-1)$,

$$K(a^{p-1} - 1) \equiv 0 \pmod{p}.$$

But K is not divisible by p, since none of its factors is; hence by the law 7), $(a^{p-1} - 1)$ must be divisible by p, i.e.

$$a^{p-1} - 1 \equiv 0 \pmod{p}.$$

This is Fermat's theorem.

To check the theorem once more, let us take $p = 23$ and $a = 5$. We then have, all modulo 23, $5^2 \equiv 2$, $5^4 \equiv 4$, $5^8 \equiv 16 \equiv -7$, $5^{16} \equiv 49 \equiv 3$, $5^{20} \equiv 12$, $5^{22} \equiv 24 \equiv 1$. With $a = 4$ instead of 5, we get, again modulo 23, $4^2 \equiv -7$, $4^3 \equiv -28 \equiv -5$, $4^4 \equiv -20 \equiv 3$, $4^8 \equiv 9$, $4^{11} \equiv -45 \equiv 1$, $4^{22} \equiv 1$.

In the example above with $a = 4$, $p = 23$, and in others, we observe that not only the $(p - 1)$st power of a, but also a smaller power may be congruent to 1. It is always true that the smallest such power, in this case 11, is a divisor of $p - 1$. (See the following Exercise 3.)

Exercises: 1) Show by similar computation that $2^8 \equiv 1 \pmod{17}$; $3^8 \equiv -1 \pmod{17}$; $3^{14} \equiv -1 \pmod{29}$; $2^{14} \equiv -1 \pmod{29}$; $4^{14} \equiv 1 \pmod{29}$; $5^{14} \equiv 1 \pmod{29}$.

2) Check Fermat's theorem for $p = 5$, 7, 11, 17, and 23 with different values of a.

3) Prove the general theorem: The smallest positive integer e for which $a^e \equiv 1 \pmod{p}$ must be a divisor of $p - 1$. (Hint: Divide $p - 1$ by e, obtaining

$$p - 1 = ke + r,$$

where $0 \leq r < e$, and use the fact that $a^{p-1} \equiv a^e \equiv 1 \pmod{p}$.)

3. Quadratic Residues

Referring to the examples for Fermat's theorem, we find that not only is $a^{p-1} \equiv 1 \pmod{p}$ always, but (if p is a prime different from 2, therefore odd and of the form $p = 2p' + 1$) that for some values of a, $a^{p'} = a^{(p-1)/2} \equiv 1 \pmod{p}$. This fact suggests a chain of interesting investigations. We may write the theorem in the following form:

$$a^{p-1} - 1 = a^{2p'} - 1 = (a^{p'} - 1)(a^{p'} + 1) \equiv 0 \pmod{p}.$$

Since a product is divisible by p only if one of the factors is, it appears immediately that either $a^{p'} - 1$ or $a^{p'} + 1$ must be divisible by p, so that for any prime $p > 2$ and any number a not divisible by p, either

$$a^{(p-1)/2} \equiv 1 \quad \text{or} \quad a^{(p-1)/2} \equiv -1 \pmod{p}.$$

From the beginning of modern number theory mathematicians have been interested in finding out for what numbers a we have the first case and for what numbers the second. Suppose a is congruent modulo p to the square of some number x,

$$a \equiv x^2 \pmod{p}.$$

Then $a^{(p-1)/2} \equiv x^{p-1}$, which according to Fermat's theorem is congruent to 1 modulo p. A number a, not a multiple of p, which is congruent modulo p to the square of some number is called a *quadratic residue of* p, while a number b, not a multiple of p, which is not congruent to any square is called a *quadratic non-residue of* p. We have just seen that every quadratic residue a of p satisfies the congruence $a^{(p-1)/2} \equiv 1 \pmod{p}$. Without serious difficulty it can be proved that for every non-residue b we have the congruence $b^{(p-1)/2} \equiv -1 \pmod{p}$. Moreover, we shall presently show that among the numbers $1, 2, 3, \cdots, p-1$ there are exactly $(p-1)/2$ quadratic residues and $(p-1)/2$ non-residues.

Although much empirical data could be gathered by direct computation, it was not easy at first to discover general laws governing the distribution of quadratic residues and non-residues. The first deep-lying property of these residues was observed by Legendre (1752–1833), and later called by Gauss the *Law of Quadratic Reciprocity*. This law concerns the behavior of two different primes p and q, and states that q is a quadratic residue of p if and only if p is a quadratic residue of q, provided that the product $\left(\frac{p-1}{2}\right)\cdot\left(\frac{q-1}{2}\right)$ is *even*. In case this product is *odd*, the situation is reversed, so that p is a residue of q if and only if q is a *non-residue* of p. One of the achievements of the young Gauss was to give the first rigorous proof of this remarkable theorem, which had long been a challenge to mathematicians. Gauss's first proof was by no means simple, and the reciprocity law is not too easy to establish even today, although a great many different proofs have been published. Its true significance has come to light only recently in connection with modern developments in algebraic number theory.

As an example illustrating the distribution of quadratic residues, let us choose $p = 7$. Then, since

$$0^2 \equiv 0, \quad 1^2 \equiv 1, \quad 2^2 \equiv 4, \quad 3^2 \equiv 2, \quad 4^2 \equiv 2, \quad 5^2 \equiv 4, \quad 6^2 \equiv 1,$$

all modulo 7, and since the remaining squares repeat this sequence, the quadratic residues of 7 are the numbers congruent to 1, 2, or 4, while the non-residues are congruent to 3, 5, or 6. In the general case, the quadratic residues of p consist of the numbers congruent to $1^2, 2^2, \cdots, (p-1)^2$. But these are congruent in pairs, for

$$x^2 \equiv (p-x)^2 \pmod{p} \qquad (\text{e.g., } 2^2 \equiv 5^2 \pmod{7}),$$

since $(p - x)^2 = p^2 - 2px + x^2 \equiv x^2 \pmod{p}$. Hence half the numbers $1, 2, \cdots, p - 1$ are quadratic residues of p and half are quadratic non-residues.

To illustrate the quadratic reciprocity law, let us choose $p = 5$, $q = 11$. Since $11 \equiv 1^2 \pmod{5}$, 11 is a quadratic residue (mod 5); since the product $[(5 - 1)/2][(11 - 1)/2]$ is even, the reciprocity law tells us that 5 is a quadratic residue (mod 11). In confirmation of this, we observe that $5 \equiv 4^2 \pmod{11}$. On the other hand, if $p = 7$, $q = 11$, the product $[(7 - 1)/2][(11 - 1)/2]$ is odd, and indeed 11 is a residue (mod 7) (since $11 \equiv 2^2 \pmod{7}$), while 7 is a non-residue (mod 11).

Exercises: 1. $6^2 = 36 \equiv 13 \pmod{23}$. Is 23 a quadratic residue (mod 13)?

2. We have seen that $x^2 \equiv (p - x)^2 \pmod{p}$. Show that these are the *only* congruences among the numbers $1^2, 2^2, 3^2, \cdots, (p - 1)^2$.

§3. PYTHAGOREAN NUMBERS AND FERMAT'S LAST THEOREM

An interesting question in number theory is connected with the Pythagorean theorem. The Greeks knew that a triangle with sides 3, 4, 5 is a right triangle. This suggests the general question: What other right triangles have sides whose lengths are integral multiples of a unit length? The Pythagorean theorem is expressed algebraically by the equation

$$a^2 + b^2 = c^2, \tag{1}$$

where a and b are the lengths of the legs of a right triangle and c is the length of the hypotenuse. The problem of finding *all* right triangles with sides of integral length is thus equivalent to the problem of finding all integer solutions (a, b, c) of equation (1). Any such triple of numbers is called a *Pythagorean number triple.*

The problem of finding all Pythagorean number triples can be solved very simply. If a, b and c form a Pythagorean number triple, so that $a^2 + b^2 = c^2$, then we put, for abbreviation, $a/c = x$, $b/c = y$. x and y are rational numbers for which $x^2 + y^2 = 1$. We then have $y^2 = (1 - x)(1 + x)$, or $y/(1 + x) = (1 - x)/y$. The common value of the two sides of this equation is a number t which is expressible as the quotient of two integers, u/v. We can now write $y = t(1 + x)$ and $(1 - x) = ty$, or

$$tx - y = -t, \qquad x + ty = 1.$$

From these simultaneous equations we find immediately that

$$x = \frac{1 - t^2}{1 + t^2}, \qquad y = \frac{2t}{1 + t^2}.$$

Substituting for x, y and t, we have

$$\frac{a}{c} = \frac{v^2 - u^2}{u^2 + v^2}, \qquad \frac{b}{c} = \frac{2uv}{u^2 + v^2}.$$

Therefore

$$\begin{aligned} a &= (v^2 - u^2)r, \\ b &= (2uv)r, \\ c &= (u^2 + v^2)r, \end{aligned} \tag{2}$$

for some rational factor of proportionality r. This shows that if (a, b, c) is a Pythagorean number triple, then a, b, c are proportional to $v^2 - u^2$, $2uv$, $u^2 + v^2$, respectively. Conversely, it is easy to see that any triple (a, b, c) defined by (2) is a Pythagorean triple, for from (2) we obtain

$$\begin{aligned} a^2 &= (u^4 - 2u^2v^2 + v^4)r^2, \\ b^2 &= (4u^2v^2)r^2, \\ c^2 &= (u^4 + 2u^2v^2 + v^4)r^2, \end{aligned}$$

so that $a^2 + b^2 = c^2$.

This result may be simplified somewhat. From any Pythagorean number triple (a, b, c) we may derive infinitely many other Pythagorean triples (sa, sb, sc) for any positive integer s. Thus, from (3, 4, 5) we obtain (6, 8, 10), (9, 12, 15), etc. Such triples are not essentially distinct, since they correspond to similar right triangles. We shall therefore define a *primitive* Pythagorean number triple to be one where a, b, and c have no common factor. It can then be shown that *the formulas*

$$\begin{aligned} a &= v^2 - u^2, \\ b &= 2uv, \\ c &= u^2 + v^2, \end{aligned}$$

for any positive integers u *and* v *with* v $>$ u, *where* u *and* v *have no common factor and are not both odd, yield all primitive Pythagorean number triples.*

**Exercise:* Prove the last statement.

As examples of primitive Pythagorean number triples we have $u = 2$, $v = 1$: (3, 4, 5), $u = 3$, $v = 2$: (5, 12, 13), $u = 4$, $v = 3$: (7, 24, 25), $\cdots$, $u = 10$, $v = 7$: (51, 140, 149), etc.

This result concerning Pythagorean numbers naturally raises the question as to whether integers a, b, c can be found for which $a^3 + b^3 = c^3$ or $a^4 + b^4 = c^4$, or, in general, whether, for a given positive integral exponent $n > 2$, the equation

$$a^n + b^n = c^n \tag{3}$$

can be solved with positive integers a, b, c. An answer was provided by Fermat in a spectacular way. Fermat had studied the work of Diophantus, the ancient contributor to number theory, and was accustomed to making comments in the margin of his copy. Although he stated many theorems there without bothering to give proofs, all of them have subsequently been proved, with but one significant exception. While commenting on Pythagorean numbers, Fermat stated that *the equation* (3) *is not solvable in integers for any* $n > 2$, but that the elegant proof which he had found was unfortunately too long for the margin in which he was writing.

Fermat's general statement has never been proved true or false, despite the efforts of some of the greatest mathematicians since his time. The theorem has indeed been proved for many values of n, in particular, for all $n < 619$, but not for all n, although no counter-example has ever been produced. Although the theorem itself is not so important mathematically, attempts to prove it have given rise to many important investigations in number theory. The problem has also aroused much interest in non-mathematical circles, due in part to a prize of 100,000 marks offered to the person who should first give a solution and held in trust at the Royal Academy at Göttingen. Until the post-war German inflation wiped out the monetary value of this prize, a great number of incorrect "solutions" was presented each year to the trustees. Even serious mathematicians sometimes deceived themselves into handing in or publishing proofs which collapsed after some superficial mistake was discovered. General interest in the question seems to have abated since the devaluation of the mark, though from time to time there is an announcement in the press that the problem has been solved by some hitherto unknown genius.

§4. THE EUCLIDEAN ALGORITHM

1. General Theory

The reader is familiar with the ordinary process of long division of one integer a by another integer b and knows that the process can be carried

out until the remainder is smaller than the divisor. Thus if $a = 648$ and $b = 7$ we have a quotient $q = 92$ and a remainder $r = 4$.

$$\begin{array}{r|l} & 92 \\ \hline 7 & 648 \\ & 63 \\ \hline & 18 \\ & 14 \\ \hline & 4 \end{array} \qquad 648 = 7 \cdot 92 + 4.$$

We may state this as a general theorem: *If* a *is any integer and* b *is any integer greater than* 0, *then we can always find an integer* q *such that*

$$a = b \cdot q + r, \tag{1}$$

where r *is an integer satisfying the inequality* $0 \leq r < b$.

To prove this statement without making use of the process of long division we need only observe that any integer a is either itself a multiple of b,

$$a = bq,$$

or lies between two successive multiples of b,

$$bq < a < b(q + 1) = bq + b.$$

In the first case the equation (1) holds with $r = 0$. In the second case we have, from the first of the inequalities above,

$$a - bq = r > 0,$$

while from the second inequality we have

$$a - bq = r < b,$$

so that $0 < r < b$ as required by (1).

From this simple fact we shall deduce a variety of important consequences. The first of these is a method for finding the greatest common divisor of two integers.

Let a and b be any two integers, not both equal to 0, and consider the set of all positive integers which divide both a and b. This set is certainly finite, since if a, for example, is $\neq 0$, then no integer greater in magnitude than a can be a divisor of a, to say nothing of b. Hence there can be but a finite number of common divisors of a and b, and of these let d be the greatest. The integer d is called the *greatest common divisor* of a and b, and written $d = (a, b)$. Thus for $a = 8$ and $b = 12$ we find by direct trial that $(8, 12) = 4$, while for $a = 5$ and $b = 9$ we find that $(5, 9) = 1$. When a and b are large, say $a = 1804$ and $b = 328$, the attempt to find (a, b) by trial and error would be quite wearisome.

A short and certain method is provided by the *Euclidean algorithm.* (An algorithm is a systematic method for computation.) It is based on the fact that from any relation of the form

$$a = b \cdot q + r \tag{2}$$

it follows that

$$(a, b) = (b, r). \tag{3}$$

For any number u which divides both a and b,

$$a = su, \qquad b = tu,$$

also divides r, since $r = a - bq = su - qtu = (s - qt)u$; and conversely, every number v which divides b and r,

$$b = s'v, \qquad r = t'v,$$

also divides a, since $a = bq + r = s'vq + t'v = (s'q + t')v$. Hence *every* common divisor of a and b is at the same time a common divisor of b and r, and conversely. Since, therefore, the set of *all* common divisors of a and b is identical with the set of all common divisors of b and r, the *greatest* common divisor of a and b must be equal to the greatest common divisor of b and r, which establishes (3). The usefulness of this relation will be seen immediately.

Let us return to the question of finding the greatest common divisor of 1804 and 328. By ordinary long division

$$\begin{array}{r} 5 \\ 328\,\overline{)\,1804} \\ \underline{1640} \\ 164 \end{array}$$

we find that

$$1804 = 5 \cdot 328 + 164.$$

Hence from (3) we conclude that

$$(1804, 328) = (328, 164).$$

Observe that the problem of finding (1804, 328) has been replaced by a problem involving smaller numbers. We may continue the process. Since

$$\begin{array}{r} 2 \\ 164\,\overline{)\,328} \\ \underline{328} \\ 0, \end{array}$$

we have $328 = 2 \cdot 164 + 0$, so that $(328, 164) = (164, 0) = 164$. Hence $(1804, 328) = (328, 164) = (164, 0) = 164$, which is the desired result.

This process for finding the greatest common divisor of two numbers is given in a geometric form in Euclid's *Elements*. For arbitrary integers a and b, not both 0, it may be described arithmetically in the following terms.

We may suppose that $b \neq 0$, since $(a, 0) = a$. Then by successive division we can write

$$
\begin{aligned}
a &= bq_1 + r_1 && (0 < r_1 < b) \\
b &= r_1 q_2 + r_2 && (0 < r_2 < r_1) \\
r_1 &= r_2 q_3 + r_3 && (0 < r_3 < r_2) \\
r_2 &= r_3 q_4 + r_4 && (0 < r_4 < r_3) \\
&\cdots\cdots\cdots\cdots && \cdots\cdots\cdots\cdots
\end{aligned} \tag{4}
$$

so long as the remainders $r_1, r_2, r_3, \cdots$ are not 0. From an inspection of the inequalities at the right, we see that the successive remainders form a steadily decreasing sequence of positive numbers:

$$b > r_1 > r_2 > r_3 > r_4 > \cdots > 0. \tag{5}$$

Hence after at most b steps (often many fewer, since the difference between two successive r's is usually greater than 1) the remainder 0 must appear:

$$
\begin{aligned}
r_{n-2} &= r_{n-1} q_n + r_n \\
r_{n-1} &= r_n q_{n+1} + 0.
\end{aligned}
$$

When this occurs we know that

$$(a, b) = r_n ;$$

in other words, (a, b) *is the last positive remainder in the sequence* (5). This follows from successive application of the equality (3) to the equations (4), since from successive lines of (4) we have

$$(a, b) = (b, r_1), \qquad (b, r_1) = (r_1, r_2), \qquad (r_1, r_2) = (r_2, r_3),$$

$$(r_2, r_3) = (r_3, r_4), \cdots, (r_{n-1}, r_n) = (r_n, 0) = r_n .$$

Exercise: Carry out the Euclidean algorithm for finding the greatest common divisor of (a) 187, 77. (b) 105, 385. (c) 245, 193.

An extremely important property of (a, b) can be derived from equations (4). *If* d = (a, b), *then positive or negative integers* k *and* l *can be*

found such that

$$d = ka + lb. \tag{6}$$

To show this, let us consider the sequence (5) of successive remainders. From the first equation in (4)

$$r_1 = a - q_1 b,$$

so that r_1 can be written in the form $k_1 a + l_1 b$ (in this case $k_1 = 1$, $l_1 = -q_1$). From the next equation,

$$\begin{aligned} r_2 &= b - q_2 r_1 = b - q_2(k_1 a + l_1 b) \\ &= (-q_2 k_1)a + (1 - q_2 l_1)b = k_2 a + l_2 b. \end{aligned}$$

Clearly this process can be repeated through the successive remainders $r_3, r_4, \cdots$ until we arrive at a representation

$$r_n = ka + lb,$$

as was to be proved.

As an example, consider the Euclidean algorithm for finding (61, 24); the greatest common divisor is 1 and the desired representation for 1 can be computed from the equations

$$61 = 2 \cdot 24 + 13, \quad 24 = 1 \cdot 13 + 11, \quad 13 = 1 \cdot 11 + 2,$$
$$11 = 5 \cdot 2 + 1, \quad 2 = 2 \cdot 1 + 0.$$

We have from the first of these equations

$$13 = 61 - 2 \cdot 24,$$

from the second,

$$11 = 24 - 13 = 24 - (61 - 2 \cdot 24) = -61 + 3 \cdot 24,$$

from the third,

$$2 = 13 - 11 = (61 - 2 \cdot 24) - (-61 + 3 \cdot 24) = 2 \cdot 61 - 5 \cdot 24,$$

and from the fourth,

$$1 = 11 - 5 \cdot 2 = (-61 + 3 \cdot 24) - 5(2 \cdot 61 - 5 \cdot 24) = -11 \cdot 61 + 28 \cdot 24.$$

2. Application to the Fundamental Theorem of Arithmetic

The fact that $d = (a, b)$ can always be written in the form $d = ka + lb$ may be used to give a proof of the fundamental theorem of arithmetic that is independent of the proof given on page 23. First we shall prove, as a lemma, the corollary of page 24, and then from this lemma we shall deduce the fundamental theorem, thus reversing the previous order of proof.

Lemma: *If a prime* p *divides a product* ab, *then* p *must divide* a *or* b.

If a prime p does not divide the integer a, then $(a, p) = 1$, since the only divisors of p are p and 1. Hence we can find integers k and l such that

$$1 = ka + lp.$$

Multiplying both sides of this equation by b we obtain

$$b = kab + lpb.$$

Now if p divides ab we can write

$$ab = pr,$$

so that

$$b = kpr + lpb = p(kr + lb).$$

from which it is evident that p divides b. Thus we have shown that if p divides ab but does not divide a then it must divide b, so that in any event p must divide a or b if it divides ab.

The extension to products of more than two integers is immediate. For example, if p divides abc, then by twice applying the lemma we can show that p must divide at least one of the integers a, b, and c. For if p divides neither a, b, nor c, then it cannot divide ab and hence cannot divide $(ab)c = abc$.

Exercise: The extension of this argument to products of any number n of integers requires the explicit or implicit use of the principle of mathematical induction. Supply the details of this argument.

From this result the fundamental theorem of arithmetic follows at once. Let us suppose given any two decompositions of a positive integer N into primes:

$$N = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s .$$

Since p_1 divides the left side of this equation, it must also divide the right, and hence, by the previous exercise, must divide one of the factors q_k. But q_k is a prime, therefore p_1 must be equal to this q_k. After these equal factors have been cancelled from the equation, it follows that p_2 must divide one of the remaining factors q_l, and hence must be equal to it. Striking out p_2 and q_l, we proceed similarly with $p_3, \cdots, p_r$. At the end of this process all the p's will be cancelled, leaving only 1 on the left side. No q can remain on the right side, since all the q's are larger than one. Hence the p's and q's will be

Every rational number can be written in the form a/b, where a and b are integers, and all these numbers can be put in an array, with a/b in the ath column and bth row. For example, 3/4 is found in the third column and fourth row of the table below. All the positive rational numbers may now be arranged according to the following scheme: in the array just defined we draw a continuous, broken line that goes through all the numbers in the array. Starting at 1, we go horizontally to the next place on the right, obtaining 2 as the second member of the sequence, then diagonally down to the left until the first column is reached at the position occupied by 1/2, then vertically down one place to 1/3, diagonally up until the first row is reached again at 3, across to 4, diagonally down to 1/4, and so on, as shown in the figure. Travelling along this broken line we arrive at a sequence 1, 2, 1/2, 1/3, 2/2, 3, 4, 3/2, 2/3, 1/4, 1/5, 2/4, 3/3, 4/2, 5, $\cdots$ containing the rational numbers in the order in which they occur along the broken line. In this sequence we now cancel all those numbers a/b for which a and b have a common factor, so that each rational number r will appear exactly once and in its simplest form. Thus we obtain a sequence

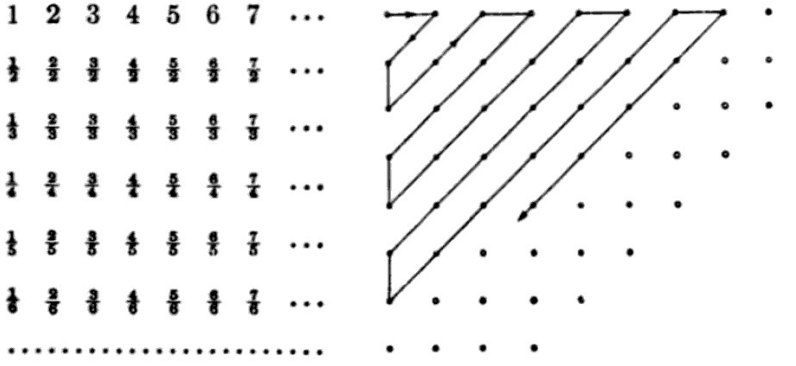


Fig. 19. Denumeration of the rational numbers.

1, 2, 1/2, 1/3, 3, 4, 3/2, 2/3, 1/4, 1/5, 5, $\cdots$ which contains each positive rational number once and only once. This shows that the set of all positive rational numbers is denumerable. In view of the fact that the rational numbers correspond in a biunique manner with the rational points on a line, we have proved at the same time that the set of positive rational points on a line is denumerable.

Exercises: 1) Show that the set of all positive and negative integers is denumerable. Show that the set of all positive and negative rational numbers is denumerable.

MATHEMATICS

"A lucid representation of the fundamental concepts and methods of the whole field of mathematics....Easily understandable." **Albert Einstein***

Written for beginners and scholars, for students and teachers, for philosophers and engineers, *What is Mathematics?* is a sparkling collection of mathematical gems that offers an entertaining and accessible portrait of the mathematical world. Brought up to date with a new chapter by Ian Stewart, this second edition offers new insights into recent mathematical developments and describes proofs of the Four-Color Theorem and Fermat's Last Theorem, problems that were still open when Courant and Robbins wrote this masterpiece, but ones that have since been solved.

A marvelously literate story, *What is Mathematics?* opens a window onto the world of mathematics.

Praise for the first edition:

"Without doubt, the work will have great influence. It should be in the hands of everyone, professional or otherwise, who is interested in scientific thinking." *The New York Times*

"A work of extraordinary perfection." *Mathematical Reviews*

"Excellent....Should prove a source of great pleasure and satisfaction." *Journal of Applied Physics*

"This book is a work of art." Marston Morse

"It is a work of high perfection....It is astonishing to what extent *What is Mathematics?* has succeeded in making clear by means of the simplest examples all the fundamental ideas and methods which we mathematicians consider the life blood of our science." Herman Weyl

The late **Richard Courant**, headed the Department of Mathematics at New York University and was Director of the Institute of Mathematical Sciences, which was subsequently renamed the Courant Institute of Mathematical Sciences. His book *Mathematical Physics* is familiar to every physicist, and his book *Differential and Integral Calculus* is acknowledged to be one of the best presentations of the subject written in modern times. **Herbert Robbins** is New Jersey Professor of Mathematical Statistics at Rutgers University. **Ian Stewart** is Professor of Mathematics at the University of Warwick, and author of *Nature's Numbers* and *Does God Play Dice?* He also writes the "Mathematical Recreations" column in *Scientific American*. In 1995 he was awarded the Royal Society's Michael Faraday medal for significant contribution to the public understanding of science.

Cover design by David Tran

Oxford Paperbacks
Oxford University Press
U.S. $21.50

9 780195 105193 90000
ISBN 0-19-510519-2