Saunders Mac Lane

Mathematics Form and Function

Springer-Verlag New York Berlin Heidelberg Tokyo

Saunders Mac Lane

Mathematics
Form and Function

With 116 Illustrations

Springer-Verlag
New York Berlin Heidelberg Tokyo

Saunders Mac Lane
Department of Mathematics
University of Chicago
Chicago, Illinois 60637
U.S.A.

AMS Classifications: 00-01, 00A05, 00A06, 03A05

Library of Congress Cataloging in Publication Data
Mac Lane, Saunders
Mathematics, form and function.
Bibliography: p.
Includes index.
1. Mathematics—1961- . I. Title.
QA39.2.M29 1986 510 85-22160

Typeset by House of Equations Inc., Newton, New Jersey.

9 8 7 6 5 4 3 2 1

ISBN-13: 978-1-4612-9340-8 e-ISBN-13: 978-1-4612-4872-9
DOI 10.1007/978-1-4612-4872-9

Contents

Introduction

This book is intended to describe the practical and conceptual origins of Mathematics and the character of its development—not in historical terms, but in intrinsic terms. Thus we ask: What is the function of Mathematics and what is its form? In order to deal effectively with this question, we must first observe what Mathematics *is*. Hence the book starts with a survey of the basic parts of Mathematics, so that the intended general questions can be answered against the background of a careful assembly of the relevant evidence. In brief, a philosophy of Mathematics is not convincing unless it is founded on an examination of Mathematics itself. Wittgenstein (and other philosophers) have failed in this regard.

The questions we endeavor to answer come in six groups, as follows.

First, what is the *Origin* of Mathematics? What are the external sources which lead to arithmetic and algebraic calculations and thence to mathematical theorems and theories? Or, are there internal sources, so that some of these theories develop just from imagination and introspection? This is close to the traditional question: Is Mathematics discovered or invented?

Second, what is the *Organization* of Mathematics? Clearly a subject so large and diverse as Mathematics requires a quite extensive and systematic organization. Traditionally, Mathematics is often split into four parts: Algebra, analysis, geometry, and applied Mathematics. This subdivision is handy at first, say for the arrangement of undergraduate courses, but it soon needs refinement. Thus number theory is to be included, perhaps as a part of algebra, but often using analysis as a tool. Finite (or "discrete") Mathematics is presently popular—but is it algebra, or logic, or applied Mathematics? Algebra soon splits into group theory, field theory, ring theory, and linear algebra (matrix theory). These split up again: number theory can be elementary, analytic, or algebraic; research in group theory is sharply divided between finite groups and infinite groups, while ring theory is split into commutative and non-commutative ring theory, with different uses and different theorems. Analysis can be real

analysis, complex analysis, or functional analysis. In geometry, algebraic geometry is based on projective geometry, differential geometry is close to parts of analysis, and topology has branches labeled point-set topology, geometric topology, differential topology, and algebraic topology. The fourth part, "applied Mathematics", is even more varied, since it may refer primarily to classical applied topics such as dynamics, fluid mechanics, and elasticity, or primarily to more recent applied topics such as systems science, game theory, statistics, operations research, or cybernetics. Finally, the active study of partial differential equations is in part applied Mathematics (especially when numerical methods are involved), in part analysis, and in part differential geometry (especially when invariant methods using differential forms are involved). But this list of subdivisions is incomplete, for example, it omits logic and foundations and their applications in computer science.

In sum, these subdivisions of Mathematics are imprecise and necessarily involve overlaps and ambiguity. The use of even finer subdivisions (as in the sixty-odd fields used by *Mathematical Reviews* to organize current research papers) still presents corresponding difficulties. Should we conclude that the real organization of Mathematics cannot be accomplished simply by subdivision into special fields? Are there deeper methods of organization? What is the proper order of the parts of Mathematics, and which branches belong first? Are there even parts of Mathematics which are unimportant or mistaken?

Since Mathematical ideas often arise in prescribed order, one may also ask whether a foundation of Mathematics provides a good organization of the subject.

Our survey will indicate that each part of Mathematics inevitably has an aspect which is formal. Factual problems necessitate calculations, but the calculations then proceed by prescriptions or by rule, rather than by continued attention to the facts of the case—yet the result of a good formal calculation does agree with the facts. Proofs in geometry flow by logical argument from axioms, but the resulting theorems fit the world. Therefore we must inquire as to the relation of the formal to the factual. Thus we begin the first chapter by exhibiting a few of the basic formal structures of Mathematics.

This leads to our third question: Are the formalisms of Mathematics based on or derived from the facts; if not, how are they derived? Alternatively, if Mathematics is a purely formal game, why do the formal conclusions fit the facts?

The fourth question is this: How does Mathematics develop? Is it motivated by quantitative questions which arise in science and engineering, is it driven by the hard problems which have arisen in the Mathematical tradition, or is it pushed by the desire to understand the tradition better? For example, how much does number theory owe to the repeated attempts to prove Fermat's last theorem? Is the solution of a famous

problem the pinnacle of Mathematical accomplishment—or should there be comparable credit to the more systematic work in the introduction of new ideas by comparison, by generalization, and by abstraction? For that matter, how does abstraction come about, and how do we know which abstraction is appropriate?

These questions about the dynamics of the development of Mathematics touch on a further—and difficult—topic: How does one evaluate the depth and importance of Mathematical research?

Careful methods and canons of proof developed first in geometry (Chapter III). Subsequently the calculus worked well, but without careful proof, using dubious notions of infinitely small quantities. This led to the problem of finding a rigorous foundation for the calculus (Chapter VI). These two cases present the fifth general question, that about rigor. Is there an absolute standard of rigor? And what are the correct foundations of Mathematics? Here there are at least six competing schools of thought, as follows.

Logicism: Bertrand Russell asserted that Mathematics is a branch of logic, and so can be founded by a development from a careful initial statement of the principles of logic. Moreover, Whitehead and Russell carried out such a development in their massive (but now neglected) book *Principia Mathematica*.

Set Theory: It is remarkable that (almost) all Mathematical objects can be constructed out of sets (and of course sets of sets). Hence arises the view that Mathematics deals just with properties of sets and that these properties can all be deduced from a suitable list of axioms for sets—either the Zermelo–Fraenkel axioms, or these axioms with supplements, some perhaps still to be discovered.

Platonism: This set-theoretic description of Mathematics is often coupled with a strong belief that these sets objectively exist in some ideal realm. Indeed some thinkers, such as Kurt Gödel, may consider that we have special means (not the usual five senses) for perceiving this ideal realm. There are other versions of platonism for Mathematics, for example one in which the ideal realms are comprised of numbers and spatial forms (the "ideal triangle").

Formalism: The Hilbert School holds that Mathematics can be regarded as a purely formal manipulation of symbols, as though in a game. This is the manipulation done when we write rigorous proofs of Mathematical theorems from axioms. This idea was part of the Hilbert program: To show that some adequate system of axioms for Mathematics is consistent, in the exact sense that proofs in the system could never lead to a contraction, such as the contradiction $0 = 1$. To this end, the proofs were to be viewed as purely formal manipulations and were to be studied objectively by strictly "finite" (and hence secure) methods. As yet, such a consistency proof has not been achieved, and Godel's famous incompleteness theorem (to be discussed in Chapter XI) makes it unlikely that it can be achieved.

Intuitionism: The Brouwer school holds that Mathematics is based on some fundamental intuitions–such as that of the sequence of natural numbers. It holds, moreover, that proofs of the existence of Mathematical objects must proceed by exhibiting these objects. For this reason intuitionism objects to some of the classical principles of logic, more explicitly the tertium non datur (either p or not p). There are a number of variants of intuitionism, some emphasizing the importance of finding proofs which are constructive.

Empiricists claim that Mathematics is a branch of empirical science, and so should have a strictly empirical foundation, say as the science of space and number.

In recent years, these (and other) standard views as to the nature and foundation of Mathematics have not been very fruitful of new insights or understanding. For this reason, we do not wish in this book to assume any one such position at the start. Instead, we intend to examine what is actually present in the practice and in the formalism of Mathematics. Only then, with the evidence before us, will we turn to the question of what is and what ought to be a foundation of Mathematics.

Our last and most fundamental question concerns the Philosophy of Mathematics. This is actually a whole bundle of questions. There are ontological questions: What are the objects of Mathematics and where do they exist (if indeed they do exist)? There are metaphysical problems: What is the nature of Mathematical truth? This is a favorite question, given that the philosophers' search for truth often will use the truths of Mathematics as the prime example of "absolute" truth. There are epistemological problems: How is it that we can have knowledge of Mathematical truth or of Mathematical objects? Here the answers may well depend on what is meant by such truth or by such an object.

There are also more immediate or more practical questions. If Mathematics is just formal or just logical deductions from axioms, how can Mathematics be so unreasonably effective in science (E. Wigner)? Put differently, why is Mathematics of such major use in understanding the world?

The various schools on the foundations have correspondingly various attempts to answer these questions, none of them generally convincing. Often–especially in work by philosophers–they are anchored almost exclusively in the most elementary parts of Mathematics–numbers and continuity. Much more substantive material is at hand. This is why we begin with a fresh view of the variety of Mathematics.

To this end, Chapter I starts with the traditional idea that Mathematics is the science of numbers and space–but shows that this starting point can lead directly to some basic formal notions (transformation group, continuity, and metric space) in defiance of the usual historical order. The next chapter describes the natural numbers as a structure, with both surface and deep aspects. The traditional foundations of geometry are sum-

marized in the third chapter, with emphasis on the ubiquitous role of groups of motions and on the remarkable observation that almost all the basic geometrical ideas can be developed in just two dimensions. The familiar (but extraordinary) fact that very many measures of magnitude (in time, space or quantity) can all be consigned to one structure—that of real numbers—is the subject of Chapter IV. The next chapter discusses the origins of the idea of "function" and the troubles in defining it. This leads through transformations to groups again and to the question: Why do the very simple group axioms lead to such deep structural results? The analysis of "effect proportional to cause" is the starting point of linear algebra (Chapter VII), but its ramifications (such as the notion of an eigenvalue) extend beyond algebra. The next chapter deals with some of the aspects of higher geometry: What is a manifold? Some of these ideas are closely tied to classical mechanics, which illustrates (Chapter IX) the intricate connection between applied and pure Mathematics. Chapter X in complex analysis returns to the study of functions—this time holomorphic functions; they are closely tied to the manifolds of Chapter VIII and to the origins of topology. At the end, the book returns to questions of foundations (Chapter XI) and then to the six philosophical questions raised above. With this sample of the extensive substance of Mathematics at hand, these questions take on a different and more illuminating form.

Our discussions of the scope of elementary Mathematics do assume some acquaintance with Mathematics; however, we endeavor to motivate and define explicitly all the Mathematical concepts which play a role in our discussion. Each defined word is italicized. A reference to §VII.6 is to the sixth section of chapter seven, while (VII.6.5) is to the fifth numbered equation of that section; references within a chapter omit the chapter number.

Since our survey touches upon many parts of classical elementary Mathematics, we assume that the reader has at hand some of his own familiar texts for possible reference. We add only occasional supplementary references to the Bibliography at the end, in the form Bourbaki [1940]. There are a number of references to *Survey* of *Modern Algebra* and to *Algebra*, both books written in some combination by Birkhoff and Mac Lane. *Homology* and *Categories Work* (short for "Categories for the Working Mathematician" refer to books by Mac Lane alone. We do note here a few other overviews of Mathematics. That magnificent multivolume monster by Bourbaki (for example, [1940]) is a splendid formal organization of many advanced topics, formulated in blissful disregard of the origins and applications which are important to our present purpose. On a more elementary level the 1977 essay by Gärding covers, with different emphasis, many of the topics on which we touch. Davis and Hersh [1981] has a more popular scope.

CHAPTER I

Origins of Formal Structure

Mathematics, at the beginning, is sometimes described as the science of Number and Space—better, of Number, Time, Space, and Motion. The need for such a science arises with the most primitive human activities. These activities presently involve counting, timing, measuring, and moving, using numbers, intervals, distances, and shapes. Facts about these operations and ideas are gradually assembled, calculations are made, until finally there develops an extensive body of knowledge, based on a few central ideas and providing formal rules for calculation. Eventually this body of knowledge is organized by a formal system of concepts, axioms, definitions, and proofs. Thus Euclid provided an axiomatization of geometry, with careful demonstrations of the theorems from the axioms; this axiomatization was perfected by Hilbert about 1900, as we will indicate in Chapter III. Similarly the natural numbers arise from counting, with notation which provides to every number the next one—its successor, and with formal rules for calculating sums and products of numbers. It then turns out that all these formal rules can be deduced from a short list of axioms (Peano–Dedekind) on the successor function (Chapter II). Finally, the measurements of time and space eventually are codified in the axioms (Chapter IV) for the real numbers. In sum, these three chapters II–IV present the standard formal axiomatization of the science of number, space, and time.

This development of the formal from the factual is a long historical process in which the leading concepts might very well have come in a different order. Our concern is not the historical order, but the very possibility of a development of form from fact. To illustrate this, we start again from number, time, space, and motion and build up directly some of the general concepts of modern Mathematics. Thus counting leads to cardinal and ordinal numbers and to infinite sets and transformations. The analysis of time leads to the notion of an ordered set and a complete ordered set; these concepts fit also with geometrical measurement. The study of motion (in space) and of the composition of two motions suggests

the notion of a transformation group. Comparison of this notion of composition with the arithmetic operations of addition and multiplication leads by further abstraction to the concept of a group. On the other hand, motion involves continuity, and the formal analysis of continuity gives rise to a simple axiomatic description of space as a metric space or, more intrinsically, as a topological space. Thus this chapter introduces the idea of the formal in terms of certain basic structures: Set, transformation, group, order, and topology. With Bourbaki, we hold that Mathematics deals with such "mother structures". Against the historical order, we hold that they arise directly from the basic stuff of Mathematics.

1. The Natural Numbers

In order to list, label, count, enumerate, or compare it is convenient to use the single system of *natural numbers*, written in our conventional decimal notation as

$$0,1,2,3,\ldots,9,10,11,\ldots. \tag{1}$$

The *same* natural numbers could be written in other notations–with the base 2 instead of 10, or as Roman numerals, or simply as marks

$$\mathrm{I},\mathrm{II},\mathrm{III},\ldots. \tag{2}$$

These numbers are used to list in order the objects of some collection of things, or simply to label these objects, or to count the collection, or to (thereby) compare two collections. From these activities, several Mathematical concepts arise together

$$\text{set}\cdot\text{number}\cdot\text{label}\cdot\text{list}.$$

At this point the word "set" simply means a collection of things: A grouping or assemblage S of objects (say, of physical objects or of symbols) such as the collection of two turtle doves, three french hens, four colley birds, or five gold rings–or the two collections

$$S = \{A,B,C\}, \qquad T = \{U,V,W\} \tag{3}$$

of three letters each, written with the conventional bracket notation for a set or collection. At this stage, the word "collection" is appropriate, because all that matters about a set (or collection) is that it is determined by specifying its elements; one does not yet need more sophisticated notions, such as sets whose elements are themselves sets, or sets of sets of sets, or sets of subsets.

In these terms, one can give semi-final descriptions of the (at first) highly informal operations of listing, labeling, counting, and comparing. To "list" a collection such as $\{A,B,C\}$ means to attach in regular order a numeral to each object in the collection; one usually begins with the numeral 1 and proceeds in order, say, as $\{A_1,B_2,C_3\}$. Note that the numerals will be adequate for this process in all cases only if there is always a next numeral; this is one origin of the idea that every natural number n has an immediate successor $s(n) = n + 1$. To "label" means to attach the same numerals to the objects of the collection, but irrespective of their order, as in $\{A_2,B_3,C_1\}$. To "count" a collection means to determine how many numerals (or which numerals) are needed to label all the objects in the collection. In this connection, note that the count, done properly, always comes out to the same answer. In particular, the numerals needed do not depend on the order in which the objects of the collections are counted: Whether it is $\{A_1,B_2,C_3\}$, $\{B_1,A_2,C_3\}$ or $\{C_1,B_2,A_3\}$, it always ends at the same 3. Comparing two collections, such as $\{A,B,C\}$ and $\{U,V,W\}$ means matching each object of the first collection with some object of the second, until both are exhausted, as in $\{A/W,\ B/V,\ C/U\}$. Of course, it might happen that one collection is exhausted before the other; the first is then "smaller" in the comparison. The result of this comparison does not depend on the order in which objects are matched: $\{A,B\}$ in any order is smaller than $\{U,V,W\}$. There are many pairs of collections to be compared, but it again turns out that it is not necessary to compare each pair; it is enough to compare finite collections with the standard initial segments of the positive natural numbers:

$$\{1,2,3\}, \qquad \{1,2,3,4,\}, \qquad \{1,2,3,4,5\}, \quad \text{etc.}$$

In this context, one says that the collection $\{A,B,C\}$ has the *cardinal* number 3, in symbols

$$\#\{A,B,C\} = 3. \tag{4}$$

As noted, this means that there is a *one-to-one correspondence* f

$$f{:}1 \mapsto A, \qquad 2 \mapsto B, \qquad 3 \mapsto C \tag{5}$$

which matches the standard collection $\{1,2,3\}$ to the collection $\{A,B,C\}$. The collection $\{U,V,W\}$ has the same cardinal number, by the correspondence

$$g{:}1 \mapsto U, \qquad 2 \mapsto V, \qquad 3 \mapsto W. \tag{6}$$

The formal definition of this matching process states that a *bijection* b (a one-to-one correspondence) from a collection S to a collection T is a rule

b which assigns to each element s in S an element $b(s)$ in T, in such a way that every element t of T occurs for exactly one s. This means that the *inverse* of b (b read backwards) is a bijection from T to S; thus the inverse of the bijection f of (5) above is

$$f^{-1}: A \mapsto 1, \qquad B \mapsto 2, \qquad c \mapsto 3. \tag{7}$$

"Composed" with the bijection g of (6) this gives a bijection, f^{-1} followed by g, directly from $\{A,B,C\}$ to $\{U,V,W\{$ as

$$g \cdot f^{-1}: A \mapsto U, \qquad B \mapsto V, \qquad C \mapsto W \tag{8}$$

Thus the elementary observation that the two collections $\{A,B,C\}$ and $\{U,V,W\}$ have the same cardinal number,

$$\#\{A,B,C\} = \#\{U,V,W\},$$

suggests the more general process of "composing" bijections, one followed by another. Indeed, these ideas about bijections can be used to provide a formal definition of the (cardinal) natural numbers (§II.8).

But, whatever the natural numbers are (or however they may be defined) their primary function is to serve in calculations of sums, products, or powers.

The *sum* of two numbers is the cardinal number one gets by combining two sets with the two given numbers, provided these sets are *disjoint*—that is, have no common elements. Thus if A, B, C, U, V above are all different, the sum $3 + 2 = 5$ is

$$3 + 2 = \#\{A,B,C,U,V\},$$

and similarly for other sums. The *product* $2 \cdot 3$ can be described "geometrically" as the cardinal number of a 2×3 square array

$$2 \cdot 3 = \# \begin{Bmatrix} (A,U)(B,U)(C,U) \\ (A,V)(B,V)(C,V) \end{Bmatrix}.$$

Here the three columns are three disjoint sets, so the product can also be described as an iterated sum

$$2 \cdot 3 = 2 + 2 + 2.$$

Similarly, the *exponential* 2^3 can be described as an iterated product

$$2^3 = 2 \cdot 2 \cdot 2;$$

it can also be described as the cardinal number of the set of all functions from a 3-element set $\{1,2,3\}$ to a 2-element set $\{0,1\}$.

These three arithmetic operations were invented (or discovered?) because they have all manner of practical uses in financial or scientific calculations. But, to make such calculations we never bother to reduce each operation to its original meaning, as this meaning has just been described. Instead, for the usual decimal notation, one may use a computer or employ the familiar rules: The addition and multiplication tables for the digits from 0 to 9, plus the rules for carry-over of tens. These rules are "formal" in the basic sense of the word: They do not refer to the meanings of the decimals or of the arithmetic operations (though they can be rigorously deduced from these meanings). Instead they simply specify what to do, and specify that correctly. Thus if one counts two disjoint collections as having 5 and 17 members, respectively, and then adds the decimals 5 and 17 according to the rules, the sum is always the count for the combined collection—and similarly for the product. To be sure, items can get lost from collections and calculators can make errors, but then there are further rules to make checks, like the rule of "casting out 9's" (replace each decimal by the sum of its digits, then add or multiply, according to the case). For numbers written in bases other than tens, there are corresponding rules for calculations and for checks (what does one cast out?).

This example gives a clear indication of what we intend to mean by *formal*: A list of rules or of axioms or of methods of proof which can be applied without attention to the "meaning" but which give results which do have the correct interpretation.

2. Infinite Sets

The collection of all the natural numbers,

$$\mathbf{N} = \{0,1,2,3,...\}, \tag{1}$$

starts with 0 and has to each number a successor; hence it is infinite. Historically, one started with 1 and not 0, but we need 0 as the cardinal number of the empty set.

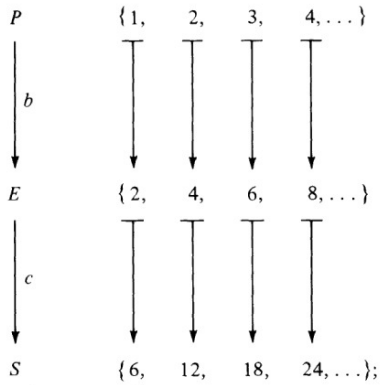
The infinite set **N** of all natural numbers includes many finite subsets

$$\{0,1,2\}, \qquad \{1,3,5,7\}, \qquad \{2,4,16\},$$

as well as infinite sets, such as the set P of all positive natural numbers

$$P = \{1,2,3,4,...\},$$

the set E of all even positive numbers, and the set S of all positive multiples of 6. These various infinite sets may be compared as follows:



(2)

the result shows that there are just as many even positives as there are positives all told; $b(n) = 2n$ defines a bijection $b: P \to E$. Similarly $c(2m) = 6m$ is a bijection $c: E \to S$. In the comparisons (2), $c(b(n)) = 6n$ gives a "composite" bijection $c \cdot b: P \to S$.

A set X is called *denumerable* when there is a bijection $f: \mathbf{N} \to X$. Thus the comparisons (2) indicate that P, E, and S are all denumerable; as a matter of fact, *any* subset of $\mathbf{N}$ is either finite or denumerable.

Two sets X and Y have the same *cardinal number* when there is a bijection $f: X \to Y$. This definition includes the finite cardinals 0,1,... already discussed in §1, and the cardinal number called $\aleph_0$ (aleph-naught) of $\mathbf{N}$, E, P, and all other denumerable sets. In this way, the elementary activity of counting leads to infinite cardinal numbers—of which $\aleph_0$ is only the first. We will later see that the set of all points on a line is infinite but not denumerable.

One can also formally describe when a set is infinite: When its cardinal number is not finite, or, equivalently, when it has a proper subset S for which there is a bijection $S \to X$.

Finitists hold that infinite sets (and geometrical infinities) are just convenient fictions, while only the finite is "real". This we must later consider. For that matter, is a finite set real? On the fourth day of Christmas, did my true love send me four colley birds or a set of four colley birds? Where is the set?

3. Permutations

A finite set, counted in any order, leads to the same (finite) cardinal number. The count is not changed by "permuting" the things counted. But one may also count how many permutations there are. Thus the set $\{1,2,3\}$ has six permutations

$$(123),\quad (231),\quad (312),\quad (213),\quad (321),\quad (132).$$

Such counts are useful in gambling or speculating. Choose three cards in succession from an (ordered) deck of thirteen; what is the chance that they come out in a direct or reverse order? It is the ratio of favorable cases [(123) or (321)] to the total number 6 of cases (of permutations). This is the root of probability, though in the end the definition of a probability must be more sophisticated than the simple ratio of favorable cases to total cases.

A permutation can be viewed "dynamically"—say, as an operation moving the original order (123) to the order (312) by the bijection

$$1 \mapsto 3, \qquad 2 \mapsto 1, \qquad 3 \mapsto 2.$$

This is usually written as a *cycle* (132), standing for $1 \mapsto 3 \mapsto 2 \mapsto 1$. Any permutation of $\{1,2,3\}$ can be viewed as a bijection

$$b\colon \{1,2,3\} \longrightarrow \{1,2,3\}$$

As a bijection, it has an inverse, and any two permutations of $\{1,2,3\}$ have a permutation as their composite.

Permutations also arise in algebra. Thus, given the polynomial

$$(x_1 + x_2)(x_3 + x_4), \tag{1}$$

what permutations of the subscripts will leave the polynomial unchanged? To begin with, one may interchange 1 and 2, or interchange 3 and 4, or do both interchanges, or do neither. These we may list as the permutations

$$(12), \quad (34), \quad (12)(34), \qquad 1; \tag{2}$$

here (12)(34) is $1 \mapsto 2$, $2 \mapsto 1$, $3 \mapsto 4$, $4 \mapsto 3$; it is the composite of the two cycles (12) and (34). Also I (do nothing) is the "identity" bijection $1 \mapsto 1$, $2 \mapsto 2$, $3 \mapsto 3$ $4 \mapsto 4$. But the polynomial (1) is also left unchanged by the following four permutations which interchange the two factors:

$$(13)(24), \quad (14)(23), \quad (1324), \quad (1423). \tag{3}$$

This completes the list. Of the 24 possible permutations of the set $\{1,2,3,4,\}$ exactly eight leave this polynomial unchanged; of these eight, four leave the factors unchanged. One may also wonder at the sequence 24, 8, 4. One may also experiment with other polynomials. Thus the polynomial

$$(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4)$$

has more symmetries (12 permutations!) while the polynomial

$$(x_1 - x_2)(x_3 - x_4) \tag{4}$$

allows only four permutations (the *four group*)

$$(12)(34), \quad (13)(24), \quad (14)(23), \quad I. \tag{5}$$

In this list, the composite of any two permutations still leaves the polynomial (4) unchanged, so the composite is also in the list. Such a list of permutations is called a *permutation group*. The combined list (2) and (3) is also such a group.

4. Time and Order

The passage of time suggests the ideas "before" and "after"; when the instant t of time comes before the instant t' we write $t < t'$. Moreover, if in turn t' is before t'', then it is apparent that t is also before t''. This can be stated formally in the *transitive law*

$$t < t' \quad \text{and} \quad t' < t'' \quad \text{imply} \quad t < t'' \tag{1}$$

for the "binary relation" $<$. Moreover, for any two distinct instants of time, one must come before. In different language, for all t and t'' exactly one of

$$t < t' \quad \text{or} \quad t = t' \quad \text{or} \quad t' < t \tag{2}$$

must hold. This statement is the law of *trichotomy*.

But the "before" and "after" of time is not the only example of these two laws. There is a "discrete" example. For natural numbers, $m < n$ means that n comes after m in the list of numbers succeeding m; here both laws (1) and (2) hold:

$$0 < 1 < 2 < 3 < \cdots . \tag{3}$$

The usual order of the positive and negative integers provides another instance of these laws:

$$\cdots \; -3 < -2 < -1 < 0 < 1 < 2 < 3 < \cdots \tag{4}$$

as does the usual ordering of the rational numbers, suggested by the display

$$\begin{aligned} \mathbf{Q}\colon\ & -\tfrac{1}{5} < \cdots < 0 \cdots < \tfrac{1}{5} \cdots < \tfrac{1}{4} \cdots \\ & < \tfrac{1}{3} < \cdots < \tfrac{2}{3} \cdots < 1 \cdots . \end{aligned} \tag{5}$$

There are numerous other examples of these two formal laws. Hence it is handy to have a name for this combined situation, as it might apply to any set X (of instants of time *or* of integers or of rationals ...).

A *binary relation* $<$ on a set X specifies that $x < y$ is true or false for any two elements x,y in X; one might also say that the relation amounts to specifying a set: The set of all those ordered pairs (x,y) with $x < y$. A *linearly ordered set* is then a set X with a binary relation $<$ for which the laws (1) and (2) hold; in other words it is a set equipped with a transitive and trichotomous relation $<$. One can then invent (or discover?) many other examples of linearly ordered sets: Finite ones such as $1 < 2 < 3 < 4$ or long infinite ones such as

$$0 < 1 < 2 < 3 < \cdots < \omega < \omega+1 < \omega+2 < \cdots, \qquad (6)$$

where ω is the first thing beyond all the finite natural numbers. (This linearly ordered set is actually the start of the infinite ordinal numbers.)

This definition is an easy first (of many) cases of a list of axioms describing a common situation with many different examples. As in other cases, the choice of axioms can vary. Thus, rather than using "before" and "after", the passage of time can be described by the notion "not later than", usually written $t \leqslant t'$. This alternative can be formalized for any linearly ordered set X. Define $x \leqslant y$ to mean $x < y$ or $x = y$. This binary relation on X is then

Transitive:	$x \leqslant y$ and $y \leqslant z$ imply $x \leqslant z$,
Reflexive:	$x \leqslant x$ for all x,
Antisymmetric:	$x \leqslant y$ and $y \leqslant x$ imply $x = y$.

Finally, corresponding to trichotomy, it has the property:

For all x and y in X, either $x \leqslant y$ or $y \leqslant x$.

Conversely, let any set X have a binary relation $\leqslant$ with these four properties, and *define* $x < y$ to mean that $x \leqslant y$ but $x \neq y$. Then X is indeed a linearly ordered set and the originally given relation $\leqslant$ is related to $<$ as before. In brief, the same notion of linear order can be defined in two formally different ways, via $<$ or via $\leqslant$. In general, the same situation may often be defined in two or more formally different ways.

One also asks when two "models" of the axioms are "essentially" the same—in the sense that the linearly ordered set of natural numbers has the same "order type" as the ordered set of even positive natural numbers:

$$2 < 4 < 6 < 8 < 10 < \cdots$$

So for linearly ordered sets X and Y an *order isomorphism* $f: X \to Y$ is defined to be a bijection of the set X on the set Y such that order is preserved: For all x_1 and x_2 in X,

$$x_1 < x_2 \qquad \text{implies } fx_1 < fx_2 . \tag{7}$$

When there is such an isomorphism f, X and Y are said to have the same *order type*. (This is like the definition of "same cardinal number" except that now one also keeps in mind the order of the elements being compared.) One can then readily prove (say) that any linearly ordered set of 4 elements is order isomorphic to the standard such set: $1 < 2 < 3 < 4$.

A general question is then at hand: Can one describe a particular model of the axioms by giving enough additional axioms to determine the model uniquely (i.e. uniquely up to an order isomorphism?) In the present case, can one give properties of an ordered set X which imply the existence of an order isomorphism $X \to \mathbf{N}$ (or $X \to \mathbf{Q}$, the ordered set of rationals, or $X \to \mathbf{R}$, the ordered set of reals?)

The answers are "yes". To get at the case of the reals $\mathbf{R}$, one must formulate the sense in which a real number (an instant of time) can be approximated by rational numbers. For example, the real number π is determined by the usual sequence of decimal approximations

$$3.14,\ 3.141,\ 3.1415,\ 3.14159,\ 3.141592,\ldots .$$

Indeed, π is the "least upper bound" of this set of rational numbers. Formally, in a linearly ordered set X an element b is an *upper bound* for a subset S of X if $s \leqslant b$ for every s in S. Also, b is a *least upper bound* for S if no b' with $b' < b$ is an upper bound for S. This implies that if S has a least upper bound, that least upper bound is unique. (This is the sense in which π, for example, is determined uniquely by its decimal expansion). Also, the set X is *unbounded* if there is in X no upper bound and no lower bound. (For example, the ordered set $\mathbf{N}$ has a lower bound 0, hence is not unbounded).

The crucial property of the ordered set of real numbers is *completeness*: Every non-empty subset S with an upper bound has a least upper bound. The additional fact that every real number can be approximated by rationals can be made formal by stating that the set $\mathbf{Q}$ of rational numbers is "dense" in $\mathbf{R}$. Here a subset D of a linearly ordered set X is said to be *dense* in X if, for all $x < y$ in X there is always a d in D between x and y, so that $x < d < y$. It is then clear that the ordered set $\mathbf{R}$ is complete, unbounded, and has a denumerable dense subset. Also one can prove that any linearly ordered set X with these three properties is order isomorphic to $\mathbf{R}$ (see Hausdorff). In the proof one uses a characterization of the order type of $\mathbf{Q}$: It is denumerable, unbounded, and dense (as a subset of itself).

This result does provide a description of the *order* of the real numbers. In Chapter IV we will combine this with a description of their algebraic properties. These properties also arise from experience with the passage of time. Once intervals of time are measured by a clock (or an hourglass) one can *add* one interval to another, and regard each instant of time t as the end of an interval (from some starting time). This addition is then an operation which produces to each pair t, t' of instants their sum, $t + t'$, with properties such as $t + t' = t' + t$ and

$$(t + t') + t'' = t + (t' + t'')$$

–just like those for the addition of natural numbers. Again, different examples lead to the same formal law.

5. Space and Motion

Space can be regarded as something extended or as a receptable for objects or as a background for ideal "figures". These aspects are all closely tied to the notion of motion through space, while motion provides the notion of measuring distance in space. Space and motion crop up together everywhere, from physics to physical exercise.

Idealization of the notion of space suggests that chunks of space are made up of figures which are filled up with "points". A point is in space, but without extent. In the extreme analysis, the space consists just of points–but to make this work the points must have added structure, say that described by giving the distance $\rho(p,q)$ from the point p to the point q. This distance is to be measured along straight lines and is a number–at first, just some rational number. But some lines must be vertical (for balance) and others horizontal. Thence comes the idea of perpendicular lines (the word suggest the vertical, as in the perpendicular version of gothic architecture). This leads to right triangles. These lead in turn to the pythagorean theorem and the discovery that the hypotheses of an isosceles right triangle with both legs of length 1 is measured by $\sqrt{2}$–which cannot be a rational number (because $\sqrt{2} = m/n$ in lowest terms would give $m^2 = 2n^2$, forcing m and then n to be even). Thus it is that space, measured with distances, requires not rational numbers but real numbers.

Thus, given the real numbers, one is led to describe space–or a chunk of space–as a collection of points $p, q, \ldots$ together with a non-negative real number $\rho(p,q)$ which is the measure of the *distance* from p to q. It is the same as the distance from q to p:

$$\rho(p,q) = \rho(q,p) \qquad \text{for all } p,q\,; \tag{1}$$

it is zero only when the points coincide:

$$\rho(p,q) \geqslant 0; \qquad \rho(p,q) = 0 \qquad \text{if and only if } p = q. \tag{2}$$

Moreover the intent is that this distance is the shortest from p to q. (The straight line is the shortest distance between two points.) In particular, this means that the distance from p to q is not lessened when it is measured along two straight lines going through a third intermediate point r. This amounts to the (Figure 1) *triangle* axiom: For all p, q and r in X,

$$\rho(p,q) \leqslant \rho(p,r) + \rho(r,q). \tag{3}$$

Thus arises the concept of a *metric space*: A collection of points p, q together with the real number distances $\rho(p,q)$ which satisfy the axioms (1), (2), and (3). The evident chunks of space—a square, a cube, a cylinder, a blob, a dumbbell, each with the usual measure of distance—are all metric spaces in the sense of this definition, as is the whole of our ("ordinary") three dimensional space. Non-Euclidean geometry (Chapter III) provides natural examples of such spaces as do the curved spaces to be considered in Chapter VIII; there are also bizarre examples—such as "a space" with infinitely many different points, with distance 1 between any two different points (try to fit *that* into the plane). Despite such bizarre examples, many elementary properties of space can be formalized and studied for a general metric space. In other words, given numbers, the Mathematical study of space need not start with the conventional ideas of Euclidean geometry, but instead with an axiom system—that of metric space—which applies to many different examples of "space".

Motion can be described in any metric space—push the points around, keeping fixed their distances apart. More formally, if F is a *figure* (a collection of points) in a metric space X a motion of F will at each time t take each point p of F to a new position (a new point) $M_t p$ in X. This passage must be "continuous" (an idea to which we will soon return). Moreover, the motion must be *rigid*—the distance apart of any two points must stay the same during the motion; in other words, for all times t and all points p and q of F, the distance ρ must satisfy

$$\rho(M_t p, M_t q) = \rho(p,q). \tag{4}$$

We speak of such a motion $(p,t) \mapsto M_t p$ as a *parametrized motion* of the figure F, with t as the parameter.

It is perhaps easier to consider just a "completed" motion—the passage from the initial position p to the final position $M_{t_1} p$ at some chosen time

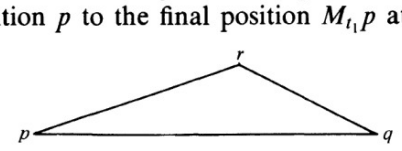


Figure 1

t_1. This is called a *rigid motion* M; it assigns to every point p of the figure concerned a new point Mp such that, for all p and q,

$$\rho(Mp,Mq) = \rho(p,q); \tag{5}$$

put briefly, a rigid motion is a bijection of space which preserves distances between points. For example, a rigid motion of an equilateral triangle into itself could be a rotation (by 120°, 240°) or a reflection of the triangle in one of the three altitudes or the identity motion (every point stays put). There are thus six such motions (symmetries) of the triangle. For motions of the plane as a whole, we will see in Chapter III the use of three typical motions: A *translation* (every line stays parallel to its original position), a *rotation* (one point is fixed) and a *reflection* (all the points on a line stay fixed. These are not all: Moving a triangle ABC into a congruent triangle $A'B'C'$ (Figure 2) may require a translation (A to A') followed by a rotation about A'; in other words, a composite motion.

From such examples arises the idea of the composition of two motions M and N–first move by M and then move the result by N, to give the *composite* motion C with

$$C(p) = N(Mp). \tag{6}$$

We write $C = N \cdot M$ for the composite and observe at once that if M and N are rigid motions, so is C. For parametrized motions the addition of time intervals usually corresponds to composites, in that

$$M_{s+t}(p) = (M_s \cdot M_t)(p). \tag{7}$$

The axioms for a metric space show that any rigid motion M keeps distinct points distinct. Indeed, $p \neq q$ implies by axiom (2) that $\rho(p,q) \neq 0$ and hence by the definition (5) of a motion that $\rho(Mp,Mq) \neq 0$), hence $Mp \neq Mq$ by axiom (2) again.

In studying the symmetry of a figure F, we usually consider a motion M of F "into" itself; that is, a motion M such that p in F moves to some $M(p)$ in F and such that every point q of F comes from some p in F, so that $q = M(p)$. By the above, the motion M is therefore a bijection of F to F, and so has an inverse M^{-1} which is also a rigid motion of F to F.

However, the reader might wish to construct an infinite figure F (say one in the plane) and a rigid motion M of F into F which is *not* onto F.

Figure 2

6. Symmetry

Symmetrical objects are all about us. There are many (man-made) symmetrical figures (Figure 1). Each of the figures has vertical symmetry, horizontal symmetry, and rotational symmetry. The vertical symmetry V can be construed as a reflection of the figure in its vertical axis, and similarly for the horizontal axis, H. The rotational symmetry can likewise be regarded as a 180° rotation R of the figure about its center. If we think of the figure as a metric space X, each of these symmetries is a rigid motion M of X onto itself, and these four motions are the only such. This suggests a definition of a *symmetry* of a figure F: A rigid motion of F onto itself. In particular the different figures of (1) have by this definition the *same* symmetry (later called the *four-group*).

By this definition, the composite of two symmetries of F is again a symmetry. Thus vertical reflection followed by another vertical reflection is the identity (which thus must count as a symmetry). Again, vertical reflection followed by horizontal reflection is the 180° rotation. This one may check by actual experiments with a rectangular card—or one may label the vertices of the rectangle by numbers 1, 2, 3, 4 so that V amounts to the permutation (12)(34), H is (14)(23), and the composite $H \cdot V$ (first apply V then H) is

$$1 \mapsto 2 \mapsto 3, \quad 2 \mapsto 1 \mapsto 4, \quad 3 \mapsto 4 \mapsto 1, \quad 4 \mapsto 3 \mapsto 2; \qquad (1)$$

this is the permutation (13)(24) given by the 180° rotation. Thus the total list of symmetries for the Figure 1 is

$$(12)(34), \quad (14)(23), \quad (13)(24), \quad I. \qquad (2)$$

This is identical to the list (3.5) of permutations allowed by the polynomial $(x_1 - x_2)(x_3 - x_4)$ of (3.4). Thus the *same* symmetry turns up in both geometric and algebraic circumstances. This suggests that *the* underlying symmetry here—in this case the "four group"—must itself be something "abstract"; neither geometric nor algebraic; or perhaps both. It need not depend on numbers—the dumbell of Figure 1 has no convenient corners to be numbered!

There are many different types of such symmetries. In three dimensions, one has the symmetry of the regular tetrahedron, or of the cube, or of the icosahedron, or of the octahedron. In the plane there are symmetrical figures such as those of Figure 2. For the equilateral triangle there are six

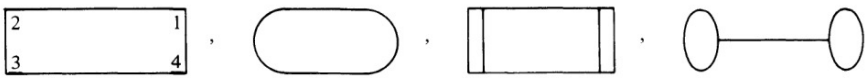


Figure 1

symmetries, accounting for all six permutations of the three vertices—or just as well, all six permutations of the three sides. For the square and also for the decorated square there are eight symmetries all told—four reflections (vertical, horizontal, and two reflections in the diagonals) and four rotations (counting the identity as a rotation through 360°!) If one labels the four vertices as in Figure 2, the eight symmetries turn out to be exactly the eight symmetries (3.2) and (3.3) listed in §3 for the polynomial $(x_1 + x_2)(x_3 + x_4)$. This again indicates that algebra and geometry have in common some underlying, more abstract, form.

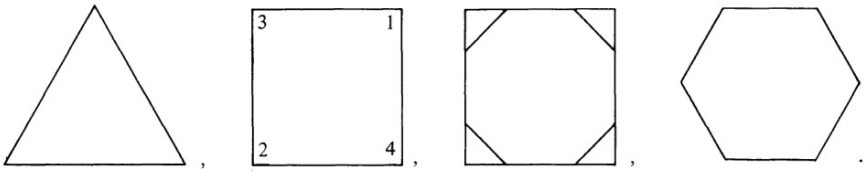


Figure 2

The frieze of a Greek temple, such as that suggested by the scheme (Figure 3) has "more" symmetry. One considers it as a "linear ornament", extending to infinity in both directions: one may picture it more schematically as in Figure 4, with nodes labeled by numbers. There are then infinitely many symmetries: Vertical reflection (n to $-n$), translation T to the right by two units and repeated such translation T^n, n times, as well as the inverse translation T^{-1} (two units left) and its iterates T^{-n}. There is also a different rigid motion S—translate one unit right *and* reflect in the horizontal axis. Then the composite $S \cdot S$ is just T, so that all the symmetries of this figure are "generated" by V and the "slide reflection" S and its inverse. If we erase the lower spikes in Figure 4 we get fewer symmetries (no S, but V and T). The reader may try to find linear ornaments with still different symmetries. (There are just seven sorts).

Three dimensional infinite symmetry comes in much greater variety. There the origin is not just from architecture, since the classification of three dimensional symmetries is the first step in the classification of crystals by the "crystallographic groups".

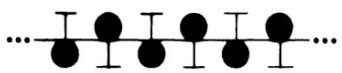


Figure 3

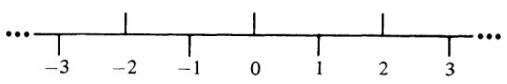


Figure 4

7. Transformation Groups

A permutation of a set, a symmetry of a figure, and a motion of Euclidean space are all examples of "transformations". A *transformation* T of a set X is a bijection $T\colon X \to X$; that is, a one-to-one correspondence $x \mapsto Tx$ on the elements x of X. Thus each transformation T has an inverse $T^{-1}\colon X \to X$; any two transformations S and T have a composite $S \cdot T$—first apply T and then S.

A *transformation group* G on a set X is a non-empty set G of transformations T on X which contains with each T its inverse and with any two transformations S, T in G their composite. This implies that G always contains the identity transformation I on X:

$$I = T \cdot T^{-1} = T^{-1} \cdot T. \tag{1}$$

A transformation group on a finite set (and especially on the typical finite set $\{1,2,\ldots,n\}$) is usually called a *permutation group*. The *symmetric group* of *degree* n is the group of all $n!$ permutations of $\{1,\ldots;n\}$.

Symmetry groups of figures or formulas are the leading examples of transformation groups, and the source of the "abstract" concept. This is a typical example of Mathematical experience leading to a formal definition. But we are also led to explicate when two transformation groups are "essentially" the same. To do this, one may examine a case such as the representation in §6 of each symmetry of the square X by a permutation of the vertices of that square. This takes place by labeling the vertices by numbers, say by a function $f\colon\{1,2,3,4\} \to X$ which puts each number on the corresponding vertex. The labeled vertices are all different; that is, $fk = fm$ implies $k = m$; one says that the function f is *injective* (an *injection*, or *one-one into*). With these labels, each motion $T\colon X \to X$ of the square sends each vertex to a vertex, so determines a permutation ${}^{\#}T\colon Y \to Y$ of the set Y of vertices. Thus ${}^{\#}T$ does to k what T does to fk; in other words,

$$f({}^{\#}T)k = T(fk), \tag{2}$$

for $k = 1,2,3$, or 4. This equation can be written in terms of composites of functions as

$$f \cdot {}^{\#}T = T \cdot f \tag{3}$$

or displayed in a diagram of the corresponding functions as

$$\begin{array}{ccc} Y & \xrightarrow{{}^{\#}T} & Y \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle f} \\ X & \xrightarrow[T]{} & X \end{array} . \tag{4}$$

This exhibits f as comparing the action of ${}^{\#}T$ on the vertices Y with the action of T on X. This diagram is called *commutative* because (3) holds: Both paths from upper left to lower right have the same result. This example (and many others like it) suggests a general formalization of the idea of comparing a transformation group H on a set Y with G on X: A *map* of (H,Y) to (G,X) is a function $f\colon Y\to X$ and a function $\#\colon G\to H$ such that (4) commutes for every transformation T in G. In case this f is an injection (as in the case above), the equation (3) shows that giving f (giving the labels of the vertices) completely determines $\#$. If moreover f is a bijection, it has an inverse f^{-1} so that $\#$ can be described directly by

$$ {}^{\#}T = f^{-1}\cdot T\cdot f\,; \tag{5}$$

to find the permutation, label each vertex by f, look to see where the vertex goes, and read off its label (by f^{-1}).

This result does formalize the evident fact that the permutations of a typical set $\{1,2,3,4\}$ of 4 things represent also the permutations of any set of four things. Generally, if sets Y and X have the same cardinal number, by a bijection $f\colon Y\to X$, then the correspondence $\#$ of (5) is a bijection from the permutation group of X to that of Y. Note incidentally that $\#$ goes in the direction opposite to f.

However, this notion of a map is a bit complicated. Moreover, it doesn't directly handle all the desired comparisons. Thus in (6.1) the dumbell Y and the perimeter X of the rectangle clearly have the "same" symmetries, but there is no evident way to get a map $f\colon Y\to X$ to make such a comparison. Indeed, there is no such f—because the dumbell Y has a center point left fixed by all the motions and there is no such point on the perimeter of the rectangle. The two transformation groups in this case can at least be compared through some intermediary—mapping each (say) into a common (containing) such rectangle.

To summarize: symmetry forces us to consider transformation groups, and even forces thoughts as to more abstractions from this notion.

8. Groups

For any three transformations R, S, and T of a set X the iterated composite, by its definition, satisfies

$$((R\cdot S)\cdot T)x = R(S(Tx)) = (R\cdot(S\cdot T))x\,,$$

so composition of transformations is associative. Now, in a transformation group G, forget the fact that the elements T of G transform things, and use only the properties of composition. It is then a group in the sense of the following definition of an "abstract" group:

A *group* is a set G equipped with three rules, as follows:

(i) A rule assigning to any two elements s, t of G on element st, called their product, such that the product is *associative*,

$$r(st) = (rs)t, \tag{1}$$

for all r, s, t in G.

(ii) A rule determining an element e (the *unit*, often written as $e = 1$) of G such that, for all t in G,

$$te = t. \tag{2}$$

(iii) A rule assigning to each t in G an element t^{-1} in G such that

$$tt^{-1} = e. \tag{3}$$

In every transformation group, composition has these properties, so every transformation group is a group. Moreover (and vice versa) Cayley's theorem asserts that every group G arises in this way from a transformation group; just take the set X of points to be transformed to be the set G itself, while each t in G is the transformation sending x in G to the product tx in G. But transformations are not the only sources of groups. With multiplication taken to be the product, the positive rational numbers or the positive real numbers or the non-zero complex numbers constitute groups. If addition is taken to be a "product", the real numbers (the instants of time) form a group, as do the ordinary clock hours ($12 = 0$). Other groups, as we will see, arise in number theory. Groups such as these, where the product is commutative,

$$st = ts \tag{4}$$

for all s and t are called *abelian groups*.

There are many consequences of the simple axioms (i), (ii), and (iii) for a group. They include easy consequences such as the cancellation law ($st = s't$ implies $s = s'$) or the rules

$$te = t = et, \qquad tt^{-1} = e = t^{-1}t \tag{5}$$

which might as well (for the sake of symmetry) be used as axioms in place of (2) and (3). A group G may have *subgroups* S (a subset which is itself a group under the same multiplication (and inverse)). If G is finite, its cardinal number is called its *order*. One proves that the order of a subgroup is always a divisor of the order of the group; this serves to understand and explain some of the observations made above about the orders 8 and 4 of subgroups of the symmetric group of four things. There are all manner of constructions of particular groups. Thus to each positive n the *cyclic group*

that always $te = t$", our axiom (ii) has specified that the element e is "given". Indeed it can be "given" as a function $e\colon \{*\} \to G$ mapping the one point set $\{*\}$ into the element e of the set G. Such a function is a *nullary operation* (on the set G). Thus the group axioms provide three operations

$$c\colon G \times G \to G, \qquad e\colon \{*\} \to G, \qquad -1\colon G \to G \tag{8}$$

a *binary operation* (multiplication), a nullary operation (unit), and a *unary operation* (inverse). These operations are required to satisfy certain identities (1), (2), and (3) which can be regarded as identities between "composites" of the initial operations (8).

Much the same pattern applies to operations of addition and multiplication (the axioms (§IV.3) for a ring or a commutative ring) and for the axioms on the algebraic operations for lattices, vector spaces, and the like.

Groups have been variously generalized. There are, for example, generalizations made by deletion of axioms. Drop the unary operation of inverse (and the axiom (iii) pertaining thereto) and one has the axioms for a *monoid*. Drop also the axiom (ii) for the unit e to get the axioms for a *semi-group*, and observe that there are various motivations for these deletions; semi-groups arise in the operation of finite state machines (the sequences of states form a semi-group) and in the composition of operations in functional analysis—but semi-groups do not have as rich a structure as do groups (How does one account for such varying richness of structures?) We will repeatedly examine generalizations by deletion.

These and many other cases illustrate the general notion of an algebraic structure: A set X with nullary, unary, binary, ternary . . . operations satisfying as axioms a variety of identities between composite operations. "Universal algebra" is concerned with the general properties of such structure. There is also a "many-sorted" universal algebra for those structures involving more than one set. A first example (two sorts) is a transformation group: A set X together with a group G of transformations on X. An even more decisive example is that of a ring R and a left module (§VII.11) over that ring. More recently, many-sorted universal algebra has proved useful in the computer science of data types.

9. Boolean Algebra

Another example of an algebra is provided by the operations such as the *intersection* and the *union* of subsets S and T of a given set X. If we write $x \in S$ for "x is an element of S" and $\iff$ for "if and only if", these operations are specified by giving the elements of the resulting subset of X as follows:

$$\text{Intersection} \quad x \in S \cap T \iff x \in S \text{ and } x \in T, \tag{1}$$

$$\text{Union} \quad x \in S \cup T \iff x \in S \text{ or } x \in T, \tag{2}$$

$$\Rightarrow \quad x \in S \Rightarrow T \iff \text{if } x \in S, \text{ then } x \in T, \tag{3}$$

$$\iff x \in T, \text{ or not } (x \in S).$$

They correspond exactly to the three propositional connectives "and", "or", and "if then". They may also be pictured by Venn diagrams; if the set X is taken to be all the points in a rectangle while S and T are respectively the points inside the ovals S and T, then two of these operations may be indicated by shaded areas as in Figure 1. There is also a unary operation, the *complement* $\neg S$ of S:

$$x \in \neg S \iff \text{not } (x \in S) \tag{4}$$

These various operations $\cap$, $\cup$, $\Rightarrow$, $\neg$ satisfy certain algebraic identities which can all be deduced from a suitable list of axioms, the axioms for *Boolean Algebra*. Thus the set $P(X)$ of all subsets of X is a Boolean algebra.

There also are operations on infinite families of sets. Thus if S_i is a subset of X for each i in some "index" set I, the (infinite) Union and intersection are defined by

$$x \in \bigcup_i S_i \iff \text{For some } i \text{ in } I, \quad x \in S_i, \tag{5}$$

$$x \in \bigcap_i S_i \iff \text{For every } i \text{ in } I, \quad x \in S_i \tag{6}$$

These operations correspond to the logical quantifiers "There exists an i" and "For all i", respectively. These connections with logic will be explored further in Chapter XI.

Boolean algebra provides a Mathematical way of representing properties, in that each property H of elements of a set X determines a subset of X; namely, the subset S consisting of all those elements which have the property

$$S = \{x \mid x \in X \text{ and } x \text{ has } H\}. \tag{7}$$

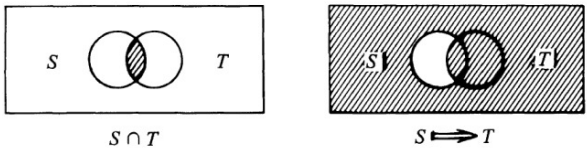


Figure 1. Boolean operations.

This subset is sometimes called the *extension* of the property H, to emphasize the notion that differently formulated properties may have the same extension–and that Mathematics has to do with extensions rather than with meanings. This in turn involves the "extensionality" axiom for sets–that a set is completely determined just by specifying its elements. This means that the equality of two subsets of X is described by the statement

$$S = T \iff (\text{For all } x \text{ in } X,\ x \in S \iff x \in T), \tag{8}$$

while the inclusion of one subset S in another is described by

$$S \subset T \iff (\text{For all } x \text{ in } X,\ x \in S \Rightarrow x \in T); \tag{9}$$

here the arrow $\Rightarrow$ stands for "implies".

This inclusion relation is transitive, reflexive, and antisymmetric, as these properties were defined in §4 above. In general, an *ordered set* W is a set W (such as $P(X)$) with a binary relation (such as $S \subset T$ for $S, T \in W$) which is transitive, reflexive, and antisymmetric. An ordered set is often said to be *partially* ordered (a *poset*) because it need not satisfy the "trichotomy" property which holds for a linear order.

It is important to recognize that many orders are just partial orders and not total orders (i.e., *not* linear). However, in many domains of the application of Mathematics to social phenomena, there is a strong tendency to order ideas, people, and institutions in a *linear* way–for example, according to rank on some imagined numerical measure. The more relevant notion of partial order seems little known and less used.

Diagrammatic presentation of an inclusion relation is suggestive. Thus the various inclusions of the subsets of a three-element set can be pictured by the rising lines in Figure 2, where the bottom symbol $\varnothing$ denotes the *empty* subset. The Boolean operations on subsets may be visualized in this figure. For example, the union $\{1,2\}$ of the subsets $\{1\}$ and $\{2\}$ is the smallest subset which lies "above" both the subsets $\{1\}$ and $\{2\}$; in this way it is the least upper bound, as defined in §4, of $\{1\}$ and $\{2\}$. Generally, the union $S \cup T$ of two subsets S and T of a set X has the properties

$$S \subset S \cup T, \qquad T \subset S \cup T, \tag{10}$$

$$S \subset R \quad \text{and} \quad T \subset R \Rightarrow S \cup T \subset R, \tag{11}$$

which state that it is the least upper bound of S and T in the partial order given by inclusion. In an exactly dual way, the intersection $S \cap T$ is the greatest lower bound of the subsets S and T. In other words, both these Boolean operations can be described directly in terms of inclusion,

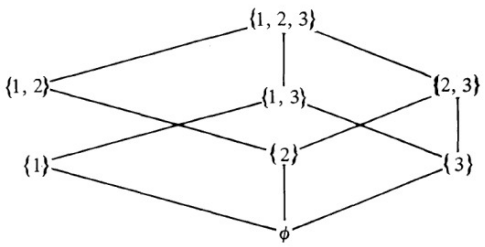


Figure 2. Lattice of subsets.

without any use of membership. In Chapter XI we will see further examples of sets treated without the use of elements.

There are corresponding definitions for other inclusion relations. In general (and in view of diagrams like that above) a poset is said to be a *lattice* when it has a top element 1, a bottom element 0 and when each pair of elements have a least upper bound (called their *join*) and a greatest lower bound (called their *meet*). The lattice of subobjects of an algebraic object is a way of describing some of the structure of that object.

10. Calculus, Continuity, and Topology

Many notions besides those of transformation groups arise from the mathematical analysis of motion. The complex motions of the planets and the varying velocities of falling bodies suggest the idea of "rate of change": Velocity as rate of change of distance or acceleration as rate of change of velocity. These ideas were codified in the notion of the derivative, subsequently formalized (Chapter VI) in the rigorous foundation of the calculus, as based on the axioms for the real numbers. This uses the definition of the derivative by means of limits and thus the consideration of a class of "good" functions—those which are differentiable. As a first example of this circle of ideas, we examine here another good class—the functions which are continuous.

A rigid motion M: $F \to F$ of a figure is continuous because (by rigidity) the distance from Mp to Mq must equal that from p to q. For a function f: $\mathbf{R} \to \mathbf{R}$ on the real numbers $\mathbf{R}$ continuity means considerably less: Just that fx and fy will be close if the originals x and y are sufficiently close. This formulation is still pretty vague; it should mean that one can make fx and fy "as close as you please" by requiring x to be "suitably close" to y. This is still vague. "As close as you please" should mean "within a specified measure δ (a positive real number) of closeness; "suitably close" should mean that one can specify a measure of closeness (again a positive real number ϵ) which will do the job. All this (and we have telescoped a

long and painful historical development) comes down to make the familiar (but meticulous) $\epsilon - \delta$ definition of continuity: A function f: $\mathbf{R} \to \mathbf{R}$ is *continuous* at a point $a \in \mathbf{R}$ if

$$\text{For all real } \epsilon > 0 \text{ there is a real } \delta > 0 \text{ such that, for all } x \text{ in } \mathbf{R}, \quad (1)$$

$$\textit{If } |x - a| < \delta, \text{ then } |f(x) - f(a)| < \epsilon. \quad (2)$$

If this statement holds for *all* points $a \in \mathbf{R}$, the function f is called continuous; the class of all such continuous functions is called C.

Note that the statement involves both propositional connectives ("if . . . then") and the so called "bounded" quantifiers (For all real numbers, there exists a real number). Thus it is that careful formulations lead to the use of concepts of formal logic.

Topological and metric spaces arise from analysis of this definition of continuity. The inequalities used in the definition arise from ideas of approximation (approximations of the value $b = f(a)$ to within the accuracy ϵ) and so implicitly involve the open interval $I_\epsilon(b) = \{y \mid |y-b| < \epsilon\}$ of center b and "radius" ϵ. In the familiar representation of the function f by its *graph* (the set of points $(x, f(x))$ in the plane), this open interval appears as an open horizontal strip of width 2ϵ around $y = f(a)$ (Figure 1). The definition is concerned with those points $x \in \mathbf{R}$ for which $f(x)$ lands in this interval $I = I_\epsilon(b)$—this set of points is usually called the *inverse image* of I under the function f, in symbols:

$$f^{-1}I = \{x \mid x \in \mathbf{R} \text{ and } f(x) \in I\}.$$

Indeed, if $x_0 \in f^{-1}I$ (that is, if $f(x_0) \in I$), then one can prove from the definition of continuity that there is an open interval (on the x axis) of center x_0 wholly contained in $f^{-1}I$. This amounts to the

Theorem. *The function f: $\mathbf{R} \to \mathbf{R}$ is continuous for all $a \in \mathbf{R}$ if and only if the inverse image $f^{-1}I$ of every open interval of $\mathbf{R}$ is a union of open intervals.*

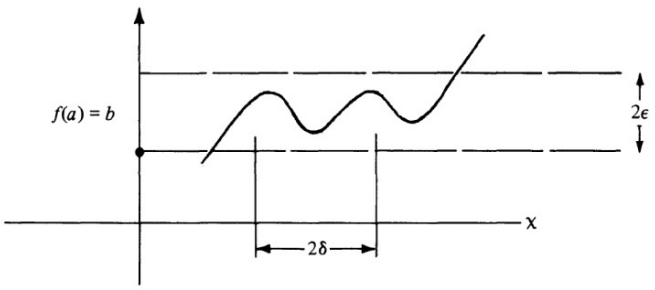


Figure 1

X and Y. The three axioms on open sets are enough to prove most of the basic facts about continuous functions—for example, the fact that the composite $x \mapsto g(f(x))$ of two continuous functions g and f is again continuous.

To describe continuity at a single point of a space, one may use the notion of "neighborhood"[11] A *neighborhood* of a point a in a topological space X is any open set of X which contains a. One then says that a function $f: X \to Y$ between topological spaces in continuous at one point $a \in X$ if to each neighborhood V of $f(a)$ there is a neighborhood U of a for which $f(U) \subset V$. This definition agrees with the previous notion of continuity at a point for a metric and expresses the intuitive idea that "nearby" points in U go into nearby points in V. Moreover, f is continuous if and only if it is continuous in this sense at each point $a \in X$.

Extensive experience has shown that this description of a "topology" in terms of open sets and neighborhoods is extraordinarily effective in formulating all sorts of Mathematical facts in a geometric form. The concept of "topology" has been appropriately abstracted from the many examples of "continuity".

The notion of a topological space was first presented by F. Hausdorff in a famous (and beautiful) book *Mengenlehre*. His definition was formulated differently, in terms of selected neighborhoods, and included an added axiom (the Hausdorff separation axiom): Two distinct points have disjoint neighborhoods. A topological space with this property is called a *Hausdorff space*.

We have now seen a number of Mathematical concepts which are described as *sets-with-structure*. Thus a linearly ordered set is a set equipped with a binary relation $<$ having certain specified properties. A group is a set equipped with a binary, a unary, and a nullary operation, which together satisfy certain identities. A Boolean algebras is similarly a set with appropriate operations. A topological space X is a set-with-structure, where in this case the "structure" consists of a specified collection of the subsets of X, namely the collection of all open sets. This kind of structure is quite different in style from the algebraic structures. There are also structures of a mixed kind. For example, there are cases of motions (e.g., translations or rotations) which deal with a set of motions which is both a group and a space. This leads to the notion of a *topological group*. Such a group is a set G which is both a group and a topological space and in which the group operations—both the product $G \times G \to G$ and the inverse $G \to G$—are continuous. It is this last condition which ties the two structures together (to make the definition complete, one must know how the topology on G induces, in a natural way, a topology on $G \times G$). As in this case, most composite axiomatic structures (combinations of two kinds of structure on the same set) involve one or more axioms expressing the formal connection between the two structures—here between the group structure and the topology.

Here is another example of a mixed structure: A *linearly ordered group* G is a set which is a group and also has a linear order, with the added axiom that $a \leqslant b$ in G and $1 \leqslant c$ implies both $ac \leqslant bc$ and $ca \leqslant cb$. This added axiom is the one which ties together the two structures of order and of multiplication. There are many examples of such linearly ordered groups–positive rational numbers or real numbers under multiplication, or integers with multiplication replaced by addition.

We will see that many Mathematical notions can be described as set-with-structure.

11. Human Activity and Ideas

This chapter, starting from the study of number, space, time, and motion, has led to the description of various formal notions–especially cardinal number, permutation, linear order, group, continuity, and topology. Each notion represents a type of formalization in Mathematics. The formalization may take the guise of a rule (e.g., a multiplication table), a simple definition (the same cardinal number), a more subtle definition (that of continuity), a list of axioms describing the common properties of several systems (linear order), a less evident such list (a group), or a list of axioms deemed sufficient to describe exactly one object (the real numbers as an ordered set). In some cases, like that of topological space, the axioms serve to help understand the common features of a wide variety of situations.

These formal notions arise largely from prematheматical concerns which can best be described as "human cultural activities". For this reason, our analysis of the genesis of Mathematics will note a number of such activities. Often it is illuminating to say that the activity leads first to a somewhat nebulous "idea", which is finally formalized, perhaps formalized in several different ways. For example, the process of counting suggests the idea of "next"–the next item to be counted or the next number to be used in the count or the next thing in some ordered list. This general idea "next" may then be formalized by a rule for adding one to each decimal or by the axioms on the operation which provide to each natural number its successor. The idea "next" appears in other forms: The next (infinite) ordinal beyond a given set of ordinals or the next step (after choice of alternative) in some computer program. Or the frequent observation of steady changes may suggest the (nebulous) idea of steady change, formalized (say) by what we called a parametrized motion.

This type of source for Mathematical form, in the cases we have noted so far, may be summarized in a table, where each activity suggests an idea and its subsequent formalizations (Table 1).

This tabulation is intended to be suggestive but not dogmatic. Each "idea" is intended to have some intuitive content; it may serve as the car-

Table 1.1

Activity	Idea	Formulation
Collecting	Collection	Set (of elements)
Counting	Next	Sucessor; order Ordinal number
Comparing	Enumeration	Bijection Cardinal number
Computing	Combination (of nos)	Rules for addition Rules for multiplication Abelian group
Rearranging	Permutation	Bijection Permutation group
Timing	Before and after	Linear order
Observing	Symmetry	Transformation group
Building, shaping	Figure; symmetry	Collection of points
Measuring	Distance; extent	Metric space
Moving	Change	Rigid motion Transformation group Rate of change
Estimating	Approximation	Continuity Limit
	Nearby	Topological space
Selecting	Part	Subset Boolean algebra
Arguing	Proof	Logical connectives
Choosing	Chance	Probability (favorable/total)
Successive actions	Followed by	Composition Transformation group

rier for the well known phenomenon of "Mathematical Intuition". The same idea may arise from different activities, and may well be the background for several different formalizations. We have tried to use familiar words to describe each idea but this does not represent any established consensus or precise definition. On the other hand, each notion, as conventionally formalized, has a rigorous definition (within some context).

The table is by no means complete; as the reader keeps it in mind he may find new examples in subsequent chapters.

Even after the basic Mathematical notions have been developed out of these activities and ideas, there continue to be inputs from outside Mathematics. These inputs often take the form of Mathematical questions arising in other sciences and requiring application of Mathematics. Thus the primitive sort of study of motion noted above becomes later the sub-

ject of dynamics (in physics) or that of celestial dynamics in astronomy. The study of social changes in part becomes the study of marginal costs or econometrics. In general, under the genesis of Mathematics we intend to include all sorts of inputs from scientific and other cultural activities.

Some formal Mathematical notions have a more complex origin. Such is the case for the notion of a "set". The idea of a collection is surely there when we count, but on this level it is hardly a useful candidate for formalization. Infinite collections also arise, perhaps at first in observations and in Euclid's proof that there are infinitely many prime numbers–but then one soon has other infinite collections. They are often subsets of (say) the set of all natural numbers, but the notion of a subset is not really forced on our attention until we try to describe the completeness of the ordered set of reals (Every bounded subset has a least upper bound) or the principle of Mathematical induction (Every set of natural numbers containing zero and the successor of each of its elements contains all the numbers). Even here we might dispense with subsets: Completeness can be described by convergent sequences and induction can be described by properties. But Boolean algebra is unthinkable without subsets. The more sophisticated notion of a set whose elements are themselves sets does arise later. The set of integers modulo 6 will be described as the set whose elements are the congruence classes such as $\{1,7,13,19, \ldots\}$, right now a topological space is most clearly defined as a set with a specified set of its subsets (namely, the open subsets). However, in both of these cases the use of sets of sets can be avoided by using relations: the relation of congruence module 6 (Gauss) or the relation stating that the subset U is a neighborhood of the point p. The real motivation for the full use of set theory lies much deeper, and will be explored in Chapter XI, where we will note the curious fact that abstract set theory arose from the study of trigonometric series!

12. Mathematical Activities

The genesis of the more complex mathematical structures tends to take place within Mathematics itself. Here there are a variety of processes which may generate new ideas and new notions. We list a few of these processes in tentative form for further refinement after our more detailed studies.

(a) *Conundrums.* Finding the solution of hard problems is one of the driving forces of Mathematical development. Fermat asserted without proof that the equations $x^n + y^n = z^n$ for $n > 2$ have no solutions in integers. As we will see in Chapter XII, this apparently innocent diophantine equation was one historical source of the whole development of algebraic number theory in the 19th century–and so was the principal origin

of such algebraic notions as that of "ideal"–although the arithmetic theory of quadratic forms also played a role.

The problem of solving polynomial equations by formulas involving radicals was a historically important conundrum. For quadratic polynomials the solution is easy, by the familiar "quadratic formula". Early algebraists found no such formulas for solutions of the general equation of 5th degree. Using permutations of the roots, it was eventually showed (by Lagrange) that such a solution was impossible–but the first real insight into the reasons for the impossibility came with Galois in 1832, (see Chapter V); this was the point where the notion of a group first explicitly arose.

Our presentation has in effect argued that the notion of a group could have arisen otherwise–but in historical perspective the solution of different Mathematical problems is a vital element in the progress of the science (and is often viewed as *the* characteristic aspect of that science).

(b) *Completion.* The whole list of natural numbers arises by starting with the first few 0, 1, 2, 3, . . . , 9 and asking that there always be a successor. Then subtraction, alas, is not always possible–until one creates all the integers. To insure the possibility of division, one must then have all the rational numbers, and so on to the real numbers and then to the complex numbers. In many other cases, the need to complete a structure under some partially defined operation brings out a new structure.

(c) *Invariance.* A non-trivial homogeneous equation

$$ax + by + cz = 0$$

has infinitely many non-zero solutions, but all can be expressed as sums of multiples of some two solutions–because, as we know, the set of all solutions (x,y,z) is a plane through the origin in 3-space and any vector lying in that plane is the sum of multiples of two suitable such. Again the solutions of the homogeneous linear second order differential equation $d^2x/dt^2 = -k^2x$ all have the form

$$x = A \cos kt + B \sin kt;$$

they are expressed here as linear combinations of two particular solutions $\cos kt$ and $\sin kt$. These two parallel situations serve to suggest the structure of a vector space (Chapter VII), the idea of a basis for such a space, and the need to describe its properties independently of any one choice of basis.

(d) *Common Structure (Analogy).* This example exhibits also the motive of finding a common structure (here, that of a vector space) underlying different but similar phenomena (here, geometry, linear equations, and linear differential equations). Another such instance is given by a description (§4) of linear order. The symmetry group as the commonality of two

striking example is the proof that a function $f\colon I \to \mathbf{R}$, continuous on a closed interval I of the reals, is uniformly continuous there. A straightforward direct proof can be given, using the basic properties of the real numbers. This proof, originally given by the German Mathematician Heine, and further developed by the French Mathematician Emil Borel, leads to the Heine–Borel theorem: If the closed interval I is the union of an (infinite) collection of opens sets U_i, so that $I = \cup U_i$, then it is a union of a finite number of these open sets,

$$I = U_{i_1} \cup \cdots \cup U_{i_k},$$

for some finite list of indices $i_1, \ldots, i_k$. In current terminology, this property states that I is a compact subset (of $\mathbf{R}$) and so leads to the idea of compactness for topological spaces.

At the end of our study of structure, we will return to a more detailed examination of these processes, internal to Mathematics, for the generation of new notions. They play a role counterpuntal to the input of problems from the sciences outside Mathematics. Both are accompanied by the continued search for deeper properties of the notions already at hand.

13. Axiomatic Structure

In the next three chapters we will indicate how number, space, and time can be described by axioms; that is, by axioms for the natural numbers, the Euclidean plane, and the real line which describe these structures uniquely. In classical terminology, these axiom systems are *categorical*, in the sense that any two "models" of the axioms, taken within an inclusive set theory, are isomorphic–as in the case described in §4 for the reals (we will also note another "first order" version of these axioms where there can be non-isomorphic non-standard models). Thus these structures are closely attached to the traditional view that (say) the axioms of Euclidean geometry describe one specific object–physical space.

In this first chapter, we have deliberately followed a different order of axiomatics, emphasizing those systems of axioms (linear order, group, metric space) which have many essentially different models. This use of axioms is historically more recent than the categorical axiomatization of geometry. In particular, it allows for the view that the formal systems studied in Mathematics come in a great variety and are intended primarily to help organize and understand selected aspects of the "real world" without being necessarily exact descriptions of a part of that unique world. For example, our presentation allows that the first step in the formalization of space could be the description of figures and chunks of space as models of metric space and not as subsets of Euclidean space. This is by no means the conventional view.

Nevertheless, this chapter has started from the conventional idea of Mathematics as the science of number, time, space, and motion, to go beyond these topics to related more general formal notions of cardinal number, permutation, order, transformation, group, and topological space. Mathematical experience, as suggested in our subsequent chapters, shows that each of these notions plays a basic role in Mathematics. We have deliberately put them first to let the reader judge their importance. This does not mean that they need be prior to the classical description of number and space, but simply that they appear in parallel to these classical notions.

The linear order of a book does not allow the actual presentation to be in parallel.

CHAPTER II

From Whole Numbers to Rational Numbers

1. Properties of Natural Numbers

Various human activities such as listing, counting, and comparing lead, as we saw, to the natural numbers

$$\mathbf{N} = \{0,1,2,3,4, \ldots\}$$

and to the operations of addition, multiplication, and exponentiation on these numbers. These operations have a variety of general properties. For example, addition for all natural numbers k, m, and n satisfies the equations

$$m + 0 = m, \qquad m + n = n + m, \tag{1}$$

$$k + (m + n) = (k + m) + n. \tag{2}$$

These rules can be proved from the definitions of the operations. For example, the *commutative* law (1) holds because, when two disjoint finite sets are combined, the cardinal number of the combined set does not depend on which of the two sets is taken first. On the other hand, the rules are *formal* in the sense that they can be used directly without attention to their "meaning". For example, the *associative* law (2) tells me that if I add a long column of figures in three successive groups, subsequently combined, the final result will be the same, irrespective of the order in which the three are combined. A similar rule will work for more than three groups. Moreover, these (long-established) rules are inviolate: If it doesn't turn out as they specify, I know that I have made a mistake somewhere. This is the merit of a formal rule: Once firmly established, it can be applied mechanically and is an infallible guide.

Multiplication has corresponding formal properties:

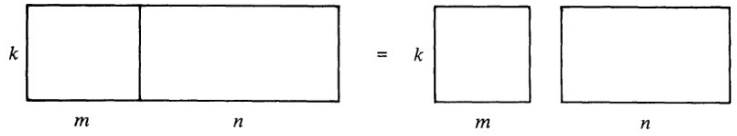
$$m \cdot 1 = m, \qquad mn = nm, \tag{3}$$

$$k(mn) = (km)n. \tag{4}$$

Together, addition and multiplication satisfy the *distributive* law

$$k(m + n) = km + kn. \tag{5}$$

Again, this law can be used formally, without attention to its origin in the definitions of addition and multiplication, as suggested in the following display:



(6)

There are many other properties of these operations. For example, every square, on division by 4, leaves a remainder 0 or 1 (never 2 or 3). If $b > 1$, then every natural number n can be written in terms of b as

$$n = a_k b^k + a_{k-1} b^{k-1} + \cdots + a_1 b + a_0 \tag{7}$$

for some natural k and with coefficients a_i all satisfying $0 \leqslant a_i < b$. In particular, if $b = 10$, this is the decimal expansion of n, and its properties lead to the familiar formal rules for manipulating decimals.

2. The Peano Postulates

Each of these properties, and many more, of the addition and multiplication of natural numbers could be demonstrated directly from the definitions of these operations on finite cardinal numbers. Such proofs would be cumbersome. However, a remarkable fact emerges: Both addition and multiplication can be described just in terms of the number zero and the single operation "add 1", and their properties can be derived from a short list of axioms on the single operation. These axioms are the Peano postulates. The idea is that the natural numbers can be listed, starting with zero, so that to each number n there is always a "next" number, its *successor* $n + 1$, and so that this process exhausts all the natural numbers. Thus we can state formally:

The (natural) numbers **N** with zero and "successors" s form a collection with the following five properties (the Peano postulates):

(i) 0 is a number;
(ii) If n is a number, so is its successor sn;
(iii) 0 is not a successor (i.e., sn is never 0);
(iv) Two numbers n, m with the same successor are equal (i.e., if $sn = sm$, then $n = m$);
(v) Let P be a property of natural numbers. If 0 has P, and if sn has P whenever n does, then P holds for all natural numbers.

This is a typical description of a structure by axioms. There are certain primitive (or undefined) terms: here the terms "number", "zero", and "successor". The statements of the axioms use only these terms and the standard logical connectives: "if . . . then", "not", "and", "equality", "for all", "there exists". Such a statement is called a formula (or a formal statement) in the language of Peano arithmetic (for more detail, see Chapter XI). In particular, a "property" of the number n, as used in postulate (v), should be one which is described by such a formula, involving n.

The induction axiom (v) is vital; it expresses the intuitive idea that taking successors exhausts all the natural numbers. It is very useful practically, in proving all sorts of formulas involving general n (for example, the formula for the sum $1 + 2^2 + 3^2 + \cdots + n^2$ of the first n squares) and for proving such results as the binomial theorem.

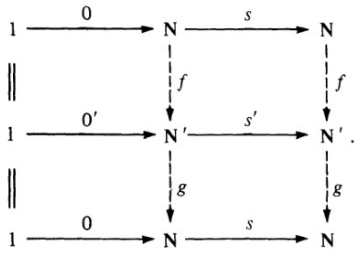
Sometimes the induction axiom is formulated in terms of sets rather than properties, as follows:

(v′) If S is a set of numbers containing 0 and if every n in S has its successor in S, then S contains all (natural) numbers.

This axiom implicitly refers to "all" subsets of **N**, so it is sometimes called a "second order" axiom, because the quantifier "all" is applied not just to elements of **N**, but also to subsets. More specifically, this form of the axiom means that we are considering the natural numbers in a context of sets, and that proofs of theorems about natural numbers may use not just the Peano axioms, but properties of sets, as these might be formulated in axioms for set theory. In this respect, it is like the completeness property of the real numbers (§I.4).

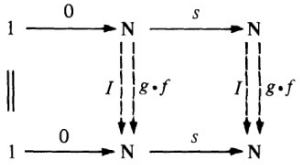
The set-theoretic induction axiom (v′) does include the property-theoretic version (v), because the usual axioms for set theory do specify that every (formal) property of elements of a set **N** does determine a subset of **N**. This transition from properties of numbers to sets of numbers is a familiar one. The use of properties may be called "intensional", because a property is described by a formula. Thus the properties "n is odd" and "n leaves the remainder 1 on division by 2" are verbally different, but describe the same set $\{1,3,5, \ldots\}$. On the other hand, as in §1.9, the use of sets is extensional: As soon as two sets include the same elements, they are equal. The "extent" of the set is all that matters.

However, the induction axiom (v) for properties is weaker than that for subsets. Since a property, as explained, can be expressed in a finite list of words in a fixed language, the number of properties of natural numbers is denumerable. However, for the usual notions of sets, a "diagonal" argument (see Chapter XI) shows that the number of subsets of **N** is not denumerable but larger. This observation has consequences. One can formulate theorems about natural numbers which are true within set theory but which cannot be proved from the Peano axioms with induction in the



(7)

Now compare the composite function $g \cdot f$: $\mathbf{N} \to \mathbf{N}$ with the identity function I: $\mathbf{N} \to \mathbf{N}$. They both make the diagram



(8)

commute. Hence, by the uniqueness assertion of (5), $g \cdot f = I$. Similarly, $f \cdot g$ is the identity function. Thus f has g as a two-sided inverse under composition, and so is a bijection.

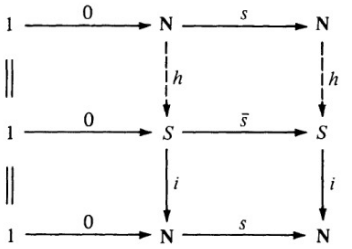
This result is typical of the axiomatic description of sets with structure. At best, such a description can determine the model only "up to isomorphism". As in this case, an isomorphism means a bijection from one model to another which "preserves" all the primitive terms involved in the axioms—as in (6) above. In this case, there are in fact many different but isomorphic models. For instance, if 100 is viewed as the zero, then the even natural numbers starting with 100 form a model for the Peano postulates when the assignment $n \mapsto n+2$ is taken to be the successor function.

3. Natural Numbers Described by Recursion

The Peano postulates are not the only possible axiomatic description of the natural numbers. Instead, one can take the recursion theorem as the (sole) axiom. In detail, this axiom assumes that the natural numbers are a set $\mathbf{N}$ with a distinguished object 0 and a function s: $\mathbf{N} \to \mathbf{N}$ which together satisfy (for all $a \in X$ and all g: $X \to X$) the recursion theorem, as pictured in the diagram (2.5). This approach to the natural numbers was first made explicit by Lawvere; it is described in some detail in (the first edition of) Mac Lane–Birkhoff.

The logical equivalence of the two approaches is readily verified. Thus, we have already seen that the Peano postulates imply the recursion axiom. Conversely, one may prove that the recursion axiom implies all the Peano postulates. The most interesting part of this demonstration is that for the

axiom of mathematical induction, for a subset S of $\mathbf{N}$, as summarized in the following diagram:



(1)

Since S is a subset of $\mathbf{N}$ each of its objects x is in $\mathbf{N}$, so the assignment $x \mapsto x$ is a function $i\colon S \to \mathbf{N}$ (the inclusion function, as displayed in the lower part of (1)). The induction assumptions on S state that 0 is in S, giving the function $0\colon 1 \to S$, and also that each n in S has its successor sn in S, so that the assignment $n \mapsto sn$ for n in S is a function $\bar{s}\colon S \to S$ as shown. By the recursion axiom, there is a function $h\colon \mathbf{N} \to S$ with $h0 = 0$, $\bar{s}h = hs$, as displayed in (1). The composite function $f = i \cdot h\colon \mathbf{N} \to \mathbf{N}$ then satisfies the same recursion conditions $f0 = 0$, $sf = fs$ as the identity function $\mathbf{N} \to \mathbf{N}$. Since our axiom asserts that the conditions determine the function uniquely, f must be the identity. Thus, each number n is $n = fn = i(hn)$, which states that n is the element hn of S and hence that the elements in S include all the elements n of $\mathbf{N}$.

This case illustrates a general point: The axioms needed to describe a Mathematical structure (here to describe the structure of $\mathbf{N}$, unique up to isomorphism) are themselves by no means unique. The recursion theorem of (2.5) is an especially convenient form of axiom; it states that the diagram $1 \to \mathbf{N} \to \mathbf{N}$ is "universal" (that is, maps uniquely into every other such diagram $1 \to X \to X$).

4. Number Theory

Once the Peano postulates are at hand, they yield all manner of specific results. Division is sometimes but not always possible, but if one tries to divide m by n one obtains a quotient q and a remainder r, which may be 0 but in any event less than n, as in the equation $m = nq + r$ with $0 \leqslant r < n$. This result is known as the *division algorithm*. Those natural numbers which have no divisors (except, of course, for themselves and 1) are the *primes*; they appear in a curious irregular order:

$$2, 3, 5, 7, 11, 13, 17, \cdots .$$

Every number n can be factored into a product of primes (some of which may be repeated). No matter how this factorization is obtained, the resulting prime factors are unique, except, of course, for their order. The proof of this unique factorization theorem rests on the division algorithm. From the prime factorization of two numbers one may read off their greatest common divisor; however, this also could be found directly from the numbers by the Euclidean algorithm, which is just an iteration of the division algorithm.

The curiously irregular sequence of primes noted above is infinite, by a proof which goes back to Euclid. One is soon led to try to estimate how thick the primes are. If $\pi(n)$ denotes the number of primes less than or equal to n, the prime number theorem (proved with more sophisticated means) will tell how fast $\pi(n)$ grows as n approaches infinity. Again, if we arrange all the numbers according to their remainder on division by 3, we get the following three arithmetic sequences

$$\begin{array}{lllll} 0 & 3 & 6 & 9 & 12, \ldots, \\ 1 & 4 & 7 & 10 & 13, \ldots, \\ 2 & 5 & 8 & 11 & 14, \ldots . \end{array}$$

Except for the prime 3 in the first sequence, all the primes must fall in the last two sequences. It turns out that there are an infinite number of primes in each of these two arithmetic sequences, and that they are, in a sense, equally distributed between those two sequences. More generally, Dirichlet's theorem asserts that any arithmetic sequence $nd + r$, for fixed d and r and increasing n, will have an infinite number of primes, provided only that d and r have no common factors except 1.

Every number can be written as a sum of at most four squares or of at most nine cubes. These results have relatively elementary proofs; by much deeper analysis for Waring's problem, similar results hold for higher powers. By trial, one can verify that each small even number can be written as a sum of two primes. Goldbach (in 1742) conjectured that this was always true. To date, no one has proved this to be so. The best results to date are Vinogradoff's: Every sufficiently large odd number r is a sum of three primes, and Chen's: Every sufficiently large even number is a sum $p + b$, where p is a prime and b is either a prime or a product of two primes.

Problems in Diophantine equations ask for solutions in integers and in natural numbers. The equation $x^2 + y^2 = z^2$ has infinitely many (well-known) solutions in non-zero integers x, y, and z, but the equation $x^4 + y^4 = z^4$ has none. Fermat stated, and no one has yet proved, that $x^n + y^n = z^n$ for $n > 2$ has none. That the numbers of such solutions is finite has just recently been proved (the Mordell conjecture). Pell's equation $x^2 - Dy^2 = 1$ has an infinite number of integer solutions, of relevance to algebraic number theory.

This is but a small sample of the wealth of questions arising for the natural numbers. All these results are ultimate consequences of the structure specified with such simplicity by the five Peano postulates.

5. Integers

To keep accounts of gains and losses, subtraction is needed. Within the natural numbers, subtraction is not always possible, but it becomes possible when the set **N** of all natural numbers is expanded to the set **Z** of integers. One can formally define the integers (and arithmetic operations upon them) in several ways. Perhaps the simplest is that of adjoining to **N** a new copy of the positive numbers, each prefixed by $-$, as $-1,-2,-3,-4,\cdots$. Then addition of the old and new integers is defined for natural numbers n and m in **N** in cases:

$$\begin{aligned} n + (-m) &= n - m, \qquad \text{if } n \geqslant m, \\ &= -(m-n), \qquad \text{if } n < m, \\ (-n) + (-m) &= -(n+m). \end{aligned}$$

With this definition, subtraction is always possible in **Z**; moreover, similar definitions describe the appropriate multiplication and order in **Z**.

Another approach to **Z** observes that subtraction amounts to solving for x an equation $n + x = m$; the ordered pair (m,n) is then introduced formally so as to denote "the" solution to this equation. The familiar rules for manipulating differences $m - n$ translate to give definitions of sum and product of such pairs by the formulas

$$(m,n)+(m',n')=(m+m', n+n'),\ (m,n)(m',n')=(mm'+nn', mn'+m').$$

But beware: the pairs (m,n) and $(m+h,n+h)$ should count as the same, hence one defines $(m,n) = (r,s)$ if and only if $m + s = n + r$, and verifies that this artificial equality satisfies the expected rules; in particular, that sums and products of equals are equal. The integers, defined to be these pairs with this equality, do not literally contain the natural numbers from which we started, but the meaning of subtraction suggests that each n in **N** be identified with the pair $(n,0)$; this identification preserves addition and multiplication. Stated more formally, this says that the function $\mathbf{N} \to \mathbf{Z}$ given by $n \mapsto (n,0)$ carries sums to sums, products to products, inequalities to inequalities, and distinct numbers to distinct integers; it is thus a monomorphism of the structure described by $+$, $\times$, and $\leqslant$.

These two constructions of the integers give essentially the same result. Specifically, the map $n \mapsto (n,0)$, $-m \mapsto (0,m)$ is an isomorphism (for

4. Hyperbolic Geometry

A set of axioms is said to be an *independent* set if no one of these axioms can be deduced from the others. It is desirable and appropriate (though not necessary) that the axioms for a basic structure, such as that of the Euclidean plane, be independent. In particular, there is the question: Is the parallel axiom independent, or can it be deduced from the others? This question has had considerable historical importance. For example, one might try to prove the parallel axiom by assuming the contrary (more than one parallel to m through a point A) and deducing a contradiction. There were several attempts to do this, most notably one in which Saccheri in 1733 deduced a large number of consequences, some of them perhaps bizarre—but none a contradiction. Nevertheless, he concluded that Euclid's parallel postulate was "vindicated". Then in the 19th century Bolyai, Lobachevsky, and Gauss took the opposite view, preparing to develop systematically a non-Euclidean geometry (specifically, a *hyperbolic* geometry) on the assumption that there is *more* than one parallel, and hence that the angle sum in a triangle is not 180°. When this is done systematically, it turns out that the angle sum is always less than 180° and that the difference between 180° and the sum is proportional to the area of the triangle.

This striking development raised (at least) two questions: "Is the resulting geometry consistent?", and "Does it fit the real world?" To answer the latter question, one must propose a specific "real world" interpretation of the primitive concepts of the geometry—say, by taking a straight line to be the path (in vacuum) of a ray of light, while an angle is the thing measured by a surveyor "turning off" with a transit the angle between two rays of light. With this interpretation, it appears that Gauss (who was also active as an astronomer) measured the angle sum for the triangle formed by chosen "points" on the peaks of three convenient mountains in Germany; the resulting angle sum was 180°, within the accuracy of the measurements then made. While the result indicates that there is not a flagrant deviation from Euclidean geometry on this interpretation, it does not provide any clear decision between the reality of Euclidean and hyperbolic geometry. It even suggests that there might never be such a decision, in view of the inevitable margin of error in the measurements made in any such interpretation. The terms involved in the interpretation are also open to question; for example, in general relativity theory the path of a light ray may not be "straight" in the intended sense. This ultimately brings up another and more profitable thought: Any geometrical axioms, Euclidean or non-Euclidean, offer a mathematical structure which may be open to a variety of different interpretations to suit a variety of geometrical (or even non-geometrical) circumstances.

There remains the question of the consistency of the assumptions of hyperbolic geometry. By definition, these assumptions are consistent if