Contents

CHAPTER 1

Before Euclid

The signature theorem of mathematics is surely the **Pythagorean theorem**, which was discovered independently in several cultures long before Euclid made it the first major theorem in his *Elements* (book 1, proposition 47). All the early roads in mathematics led to the Pythagorean theorem, no doubt because it reflects both sides of basic mathematics: number and space, or arithmetic and geometry, or the discrete and the continuous.

The arithmetic side of the Pythagorean theorem was observed in remarkable depth as early as 1800 BCE, when Babylonian mathematicians found many triples $\langle a, b, c\rangle$ of natural numbers such that $a^2 + b^2 = c^2$. Whether they viewed each triple a, b, c as sides of a right-angled triangle has been questioned; however, the connection was not missed in ancient India and China, where there were also geometric demonstrations of particular cases of the theorem.

Nevertheless, the Pythagoreans are rightly associated with the theorem because of their discovery that $\sqrt{2}$, the hypotenuse of the triangle with unit sides, is **irrational**. This discovery was a turning point in Greek mathematics, even a "crisis of foundations," because it forced a reckoning with *infinity* and, with it, the need for *proof*. In India and China, where irrationality was overlooked, there was no "crisis," hence no perceived need to develop mathematics in a deductive manner from self-evident axioms.

The nature of irrational numbers, as we will see, is a deep problem that has stimulated mathematicians for millennia. Even in antiquity, with Eudoxus's theory of proportions, the Greeks took the first step from the discrete toward the continuous.

1.1 THE PYTHAGOREAN THEOREM

For many people, the Pythagorean theorem is where geometry begins, and it is where proof begins too. Figure 1.1 shows the pure geometric form of the theorem: for a right-angled triangle (white), the square on the hypotenuse (gray) is equal to the sum of the squares on the other two sides (black).

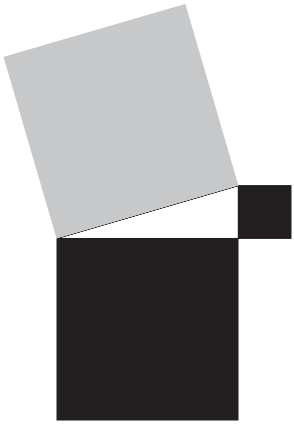

Figure 1.1 : The Pythagorean theorem

What "equality" and "sum" mean in this context can be explained immediately with the help of figure 1.2. Each half of the picture shows a large square with four copies of the triangle inside it. On the left, the large square minus the four triangles is identical with the square on the hypotenuse. On the right, the large square minus four triangles is identical with the squares on the other two sides. Therefore, the square on the hypotenuse *equals* the sum of the squares on the other two sides.

Thus we are implicitly assuming some "common notions," as Euclid called them:

1. Identical figures are equal.
2. Things equal to the same thing are equal to each other.
3. If equals are added to equals the sums are equal.
4. If equals are subtracted from equals the differences are equal.

These assumptions sound a little like algebra, and they are obviously true for numbers, but here they are being applied to geometric objects.

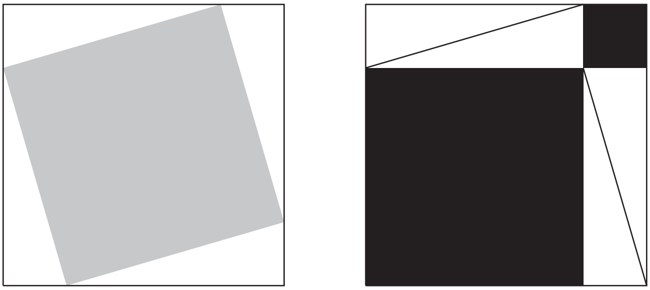

Figure 1.2 : Seeing the Pythagorean theorem

In that sense we have a purely geometric proof of a geometric theorem. The reasons why the Pythagoreans wanted to keep geometry pure will emerge in section 1.3 below.

Although figure 1.2 is as convincing as a picture can be, some might quibble that we have not really explained why the gray and black regions are squares. The Greeks who came after Pythagoras did indeed quibble about details like this, due to concerns about the nature of geometric objects that will also emerge in section 1.3. The result was Euclid's *Elements*, produced around 300 BCE, a system of proof that placed geometry on a firm (but wordy) logical foundation. Chapter 2 expands figure 1.2 into a proof in the style of Euclid. We will see that the saying "a picture is worth a thousand words" is pretty close to the mark.

Origins of the Pythagorean Theorem

As noted above, the Pythagorean theorem was discovered independently in several ancient cultures, probably earlier than Pythagoras himself. Special cases of it occur in ancient India and China, and perhaps earliest of all in Babylonia (part of modern Iraq). Thus the theorem is a fine example of the universality of mathematics. As we will see in later chapters, it recurs in different guises throughout the history of geometry, and also in number theory.

It is not known how it was first proved. The proof above is one suggestion, given by Heath (1925, 1:354) in his edition of the *Elements*. The Chinese and Indian mathematicians were more interested in triangles whose sides had particular numerical values, such as 3, 4, 5 or 5, 12, 13.

As we will see in the next section, the Babylonians developed the theory of numerical right-angled triangles to an extraordinarily high level.

1.2 PYTHAGOREAN TRIPLES

If the sides of a right-angled triangle are a, b, c, with c the hypotenuse, then the Pythagorean theorem is expressed by the equation

$$a^2 + b^2 = c^2,$$

in the algebraic notation of today. Indeed, we call a^2 "a squared" in memory of the fact that a^2 represents a square of side a. We also understand that a^2 is found by multiplying a by itself, and the Pythagoreans would have agreed with us when a is a whole number. What made the Pythagorean theorem interesting to them are the whole-number triples $\langle a, b, c\rangle$ satisfying the equation above. Today, such triples are known as **Pythagorean triples**. The simplest example is of course $\langle 3, 4, 5\rangle$, because

$$3^2 + 4^2 = 9 + 16 = 25 = 5^2,$$

but there are infinitely many Pythagorean triples. In fact, the right-angled triangles whose sides are Pythagorean triples come in infinitely many shapes because the slopes b/a of their hypotenuses can take infinitely many values.

The most impressive evidence for this fact appears on a Babylonian clay tablet from around 1800 BCE. The tablet, known as Plimpton 322 (its catalog number in a collection at Columbia University), contains columns of numbers that Neugebauer and Sachs (1945) interpreted as values of b and c in a table of Pythagorean triples. Part of the tablet is broken off, so what remains are pairs $\langle b, c\rangle$ rather than triples. Some have questioned whether the Babylonian compiler of the tablet really had right-angled triangles in mind. In my opinion, yes, because all the values $c^2 - b^2$ are perfect squares *and* the pairs $\langle b, c\rangle$ are listed in order of the values b/a—the slopes of the corresponding hypotenuses. Figure 1.3 is a completed table that includes the values of a and b/a and also a fraction x that I explain below.

The column of a values reveals something else interesting. These values are all divisible only by powers of 2, 3, and 5, which makes them particularly "round" numbers in the Babylonian system, which was based on the number 60 (some of their system survives today, with 60 minutes in a hour and 60 seconds in a minute).

We do not know how the Babylonians discovered these triples. However, the amazingly complex values of b and c can be generated from the

a	b	c	b/a	x
120	119	169	0.9917	12/5
3456	3367	4825	0.9742	64/27
4800	4601	6649	0.9585	75/32
13500	12709	18541	0.9414	125/54
72	65	97	0.9028	9/4
360	319	481	0.8861	20/9
2700	2291	3541	0.8485	54/25
960	799	1249	0.8323	32/15
600	481	769	0.8017	25/12
6480	4961	8161	0.7656	81/40
60	45	75	0.7500	2
2400	1679	2929	0.6996	48/25
240	161	289	0.6708	15/8
2700	1771	3229	0.6559	50/27
90	56	106	0.6222	9/5

Figure 1.3 : Pythagorean triples in Plimpton 322

fractions x, which are fairly simple combinations of powers of 2, 3, and 5. In terms of x, the whole numbers a, b, and c are denominator and numerators of the fractions

$$\frac{b}{a}=\frac{1}{2}\left(x-\frac{1}{x}\right) \quad \text{and} \quad \frac{c}{a}=\frac{1}{2}\left(x+\frac{1}{x}\right).$$

For example, with $x=12/5$ we get

$$\frac{1}{2}\left(x-\frac{1}{x}\right)=\frac{1}{2}\left(\frac{12}{5}-\frac{5}{12}\right)=\frac{119}{120} \quad \text{and} \quad \frac{1}{2}\left(x+\frac{1}{x}\right)=\frac{1}{2}\left(\frac{12}{5}+\frac{5}{12}\right)=\frac{169}{120}.$$

The huge triple $\langle 13500, 12709, 18541\rangle$ is similarly generated from the fraction $125/54=5^3/2\cdot 3^3$, which has roughly the same complexity as $13500=2^2\cdot 3^3\cdot 5^3$. Thus, it is plausible that the Babylonians could have generated complex Pythagorean triples by relatively simple arithmetic. At the same time, the link with geometry is hard to deny when the triples are seen to be arranged in order of the slopes b/a—an order that could not be guessed from the arrangement of a, b, c, or x values! And when one sees that these slopes cover a range of angles, roughly equally spaced, between $30°$ and $45°$ (figure 1.4), it looks as though the Babylonians were collecting triangles of different shapes.

It is also conspicuous which shape is *missing* from this collection of triangles: the one with equal sides a and b, shown in red in figure 1.4.

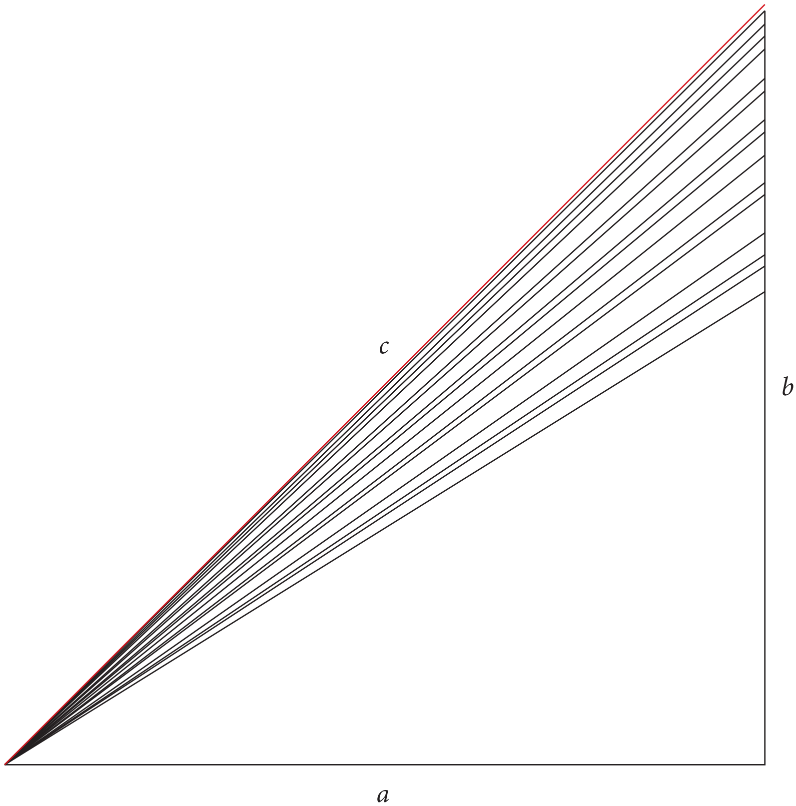

Figure 1.4 : Slopes derived from Plimpton 322

As we now know, because the Pythagoreans discovered it, this shape is missing because the hypotenuse of this triangle is *irrational*.

1.3 IRRATIONALITY

Irrationality follows naturally from the Pythagorean theorem, but apparently it was found by the Pythagoreans alone. Like other discoverers of the theorem, the Pythagoreans knew special cases with whole-number values of a, b, c. But, apparently they were the only ones to ask, Why do we find no such triples with $a = b$? The question points to its own answer: *it is contradictory to suppose there are whole numbers a and c such that* $c^2 = 2a^2$.

The argument of the Pythagoreans is not known, but the result must have been common knowledge by the time of Aristotle (384–322 BCE), as he apparently assumes his readers will understand the following brief hint:

> The diagonal of the square is incommensurable with the side, because odd numbers are equal to evens if it is supposed commensurable.
>
> (Aristotle, *Prior Analytics*, bk. 1, chap. 23)

Here "commensurable" means being a whole number multiple of a common unit of measure, so we are supposing that $c^2 = 2a^2$, where the side of the square is a units and its diagonal is c units. We reach the contradiction "odd = even" as follows.

First, by choosing the unit of measure as large as possible, we can assume that the whole numbers c and a have no common divisor (except 1). In particular, at most one of them can be even.

Now $c^2 = 2a^2$ implies that the number c^2 is even. Since the square of an odd number is odd, c must also be even, say $c = 2d$. Substituting $2d$ for c gives

$$(2d)^2 = 2a^2 \quad \text{so} \quad 2d^2 = a^2.$$

But then a similar argument shows a is even, which is a contradiction.

So it is wrong to suppose there are whole numbers a and c with $c^2 = 2a^2$.

The usual way to express this fact today is that *there are no natural numbers c and a such that* $\sqrt{2} = c/a$ or, more simply, that $\sqrt{2}$ *is irrational.*

1.4 FROM IRRATIONALS TO INFINITY

The argument for irrationality of $\sqrt{2}$ is very short and transparent in modern algebraic symbolism. Judging by the excerpt from Aristotle, it was also comprehensible enough when equations were written out in words, as the ancient Greeks did.

But there was also a geometric approach to incommensurable quantities that the Greeks called *anthyphaeresis*. It gives a different and deeper insight into the nature of $\sqrt{2}$ and, indeed, a different proof that it is irrational. Anthyphaeresis is a process that can be applied to two quantities, such as lengths or natural numbers, by repeatedly subtracting the smaller from the larger. Since it was later used to great effect by Euclid, it is today called the **Euclidean algorithm**.

More formally, given two quantities a_1 and b_1 with $a_1 > b_1$, one forms the new pair of quantities b_1 and $a_1 - b_1$ and calls the greater of them a_2

and the lesser b_2. Then one does the same with the pair a_2, b_2, and so on. For example, if $a_1 = 5$, $b_1 = 3$ we get

$$\begin{aligned}\langle a_1, b_1\rangle &= \langle 5, 3\rangle\\ \langle a_2, b_2\rangle &= \langle 3, 2\rangle\\ \langle a_3, b_3\rangle &= \langle 2, 1\rangle\\ \langle a_4, b_4\rangle &= \langle 1, 1\rangle,\end{aligned}$$

at which point the algorithm terminates because $a_4 = b_4$. The Euclidean algorithm always terminates when a_1 and b_1 are natural numbers, because subtraction produces smaller natural numbers and natural numbers cannot decrease forever. Conversely, *a ratio for which the Euclidean algorithm runs forever is irrational.*

In section 2.6 we will see the consequences of the Euclidean algorithm for natural numbers, but for the Greeks before Euclid the process of anthyphaeresis was most revealing for pairs of incommensurable quantities, such as $a_1 = \sqrt{2}$ and $b_1 = 1$. In this case the numbers a_n, b_n can and do decrease forever. In fact, we have

$$\begin{aligned}\langle a_1, b_1\rangle &= \langle \sqrt{2}, 1\rangle\\ \langle a_2, b_2\rangle &= \langle 1, \sqrt{2}-1\rangle\\ \langle a_3, b_3\rangle &= \langle 2-\sqrt{2}, \sqrt{2}-1\rangle = \langle(\sqrt{2}-1)\sqrt{2}, (\sqrt{2}-1)1\rangle,\end{aligned}$$

so $\langle a_3, b_3\rangle$ is the same as $\langle a_1, b_1\rangle$, just scaled down by the factor $\sqrt{2}-1$. Two more steps will give $\langle a_5, b_5\rangle$, again the same as $\langle a_1, b_1\rangle$ but scaled down by the factor $(\sqrt{2}-1)^2$, and so on. Thus the numbers $\langle a_n, b_n\rangle$ decrease forever, but they return to the same ratio every other step.

Since this cannot happen for any pair $\langle a, b\rangle$ of natural numbers, it follows that $\sqrt{2}$ and 1 are not in a natural number ratio; that is, $\sqrt{2}$ is irrational. Moreover, we have discovered that the pair $\langle\sqrt{2}, 1\rangle$ behaves *periodically* under anthyphaeresis, producing pairs in the same ratio every other step. It turns out, though this was not understood until algebra was better developed, that periodicity is a special phenomenon occurring with square roots of natural numbers.

Visual Form of the Euclidean Algorithm

If a and b are lengths, we can represent the pair $\{a, b\}$ by the rectangle with adjacent sides a and b. If, say, $a > b$, then the pair $\{b, a-b\}$ is represented by the rectangle obtained by cutting a square of side b from the original rectangle, shown in light gray in figure 1.5. The algorithm then repeats the process of cutting off a square in the light gray rectangle, and so on.

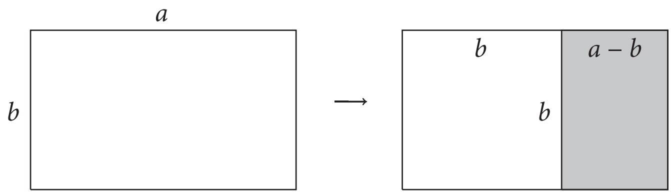

Figure 1.5 : First step of the Euclidean algorithm

When $a = \sqrt{2}$ and $b = 1$, two steps of the algorithm give the light gray rectangle shown in figure 1.6, which is the *same shape* as the original rectangle. This is because its sides are again in the ratio $\sqrt{2} : 1$, as we saw in the calculation above. Since the new rectangle is the same shape as the old, it is clear that the process of cutting off a square will continue forever.

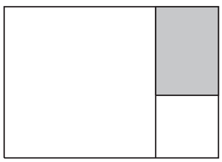

Figure 1.6 : After two steps of the algorithm on $\sqrt{2}$ and 1

The Greeks were fascinated by geometric constructions in which the original figure reappears at a reduced size. The simplest example is the so-called *golden rectangle* (see figure 1.7), in which removal of a square leaves a rectangle the same shape as the original. It follows that the Euclidean algorithm runs forever on the sides a and b of the golden rectangle, and hence these sides are in irrational ratio. This particular ratio is called the **golden ratio**.

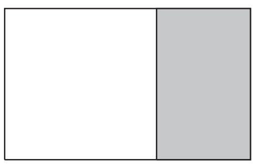

Figure 1.7 : The golden rectangle

The golden ratio is also the ratio of the diagonal to the side of the regular pentagon, where the recurrence of the original figure at reduced size can be seen in figure 1.8.

It is believed that the study of the golden ratio and the regular pentagon may go back to the Pythagoreans, in which case they were probably aware of the irrationality of the golden ratio as well as that of $\sqrt{2}$.

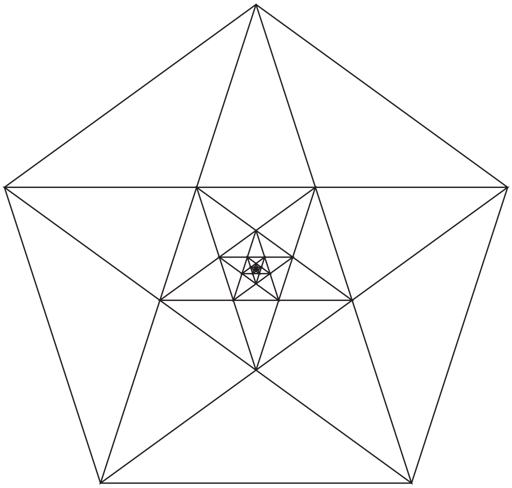

Figure 1.8 : Infinite series of pentagrams

1.5 FEAR OF INFINITY

As we have just seen, irrationality brings infinite processes to the attention of mathematicians, albeit processes of a simple and repetitive kind. At an even more primitive level, the natural numbers $0, 1, 2, 3, \ldots$ themselves represent the kind of infinity where a simple process—in this case, adding 1—is repeated without end. An infinity that involves endless repetition was called by the Greeks a *potential infinity*. They contrasted it with *actual infinity*—a somehow completed infinite totality—which was considered unacceptable or downright contradictory.

The legendary opponent of infinity was Zeno of Elea, who lived around 450 BCE. Zeno posed certain "paradoxes of the infinite," which we know only from Aristotle, who described the paradoxes only to debunk them, so we do not really know what Zeno meant by them or how they were originally stated. It will become clear, however, that Zeno accepted potential infinity while rejecting actual infinity.

A typical Zeno paradox is his first, the *paradox of the dichotomy*, in which he argues that motion is impossible because

> before any distance can be traversed, half the distance must be traversed [and so on], that these half distances are infinite in number, and that it is impossible to traverse distances infinite in number. (Aristotle, *Physics*, bk. 8, chap. 8, 263a)

Apparently, Zeno is arguing that the infinite sequence of events

reaching 1/2 way
reaching 1/4 way
reaching 1/8 way
...

cannot be completed. Aristotle answers, a few lines below this statement, that

> the element of infinity is present in the time no less than in the distance.

In other words, if one can conceive an infinite sequence of places

1/2 way, 1/4 way, 1/8 way, ...

then one can conceive an infinite sequence of times at which

1/2 way is reached, 1/4 way is reached, 1/8 way is reached, ...

Thus if Zeno is willing to admit the potential infinity of places, he has to admit the potential infinity of times. It is not a question of *completing* an infinity but only of correlating one potential infinity with another. We claim only that each of the places can be reached at a certain time; we do not have to consider the totality of places or the totality of times.

At any rate, after Zeno, Greek mathematicians handled questions about infinity by this style of argument—dealing with members of a potential infinity one by one rather than in their totality. The “actual infinity scare” was nevertheless productive, because it led to a very subtle understanding of the relation between the continuous and the discrete.

1.6 EUDOXUS

Eudoxus of Cnidus, who lived from approximately 390 BCE to 330 BCE, was a student of Plato and is believed to have taught Aristotle. His most important accomplishments are the **theory of proportions** and the **method of exhaustion**. Together, they form the summit of the Greek treatment of infinity, and they come down to us mainly through the exposition in book 5 of Euclid's *Elements*. In particular, the theory of proportions was the best treatment of rational and irrational quantities available until the nineteenth century. Indeed, it is probably the best treatment possible as long as one rejects actual infinity, which most mathematicians did until the 1870s.

The theory of proportions deals with "magnitudes" (typically lengths) and their relation to "numbers," which are natural numbers. It thereby builds a bridge between the two worlds separated by the Pythagoreans: the world of magnitudes, which vary *continuously*, and the world of counting, where numbers jump *discretely* from each number to its successor.

The theory is complicated somewhat because the Greeks thought in terms of ratios of magnitudes and ratios of numbers, without having the algebraic machinery of fractions that makes ratios easy to handle. We can understand the ratio of natural numbers m and n as the fraction m/n, so we will write the ratio of lengths a and b as the fraction a/b.[1] The key idea of Eudoxus is that ratios of lengths, a/b and c/d, are equal if and only if, for *each* natural number ratio m/n,

$$\frac{m}{n}<\frac{a}{b} \quad \text{if and only if} \quad \frac{m}{n}<\frac{c}{d}.$$

Equivalently (and this is how Eudoxus put it), for each natural number pair m and n,

$$mb<na \quad \text{if and only if} \quad md<nc.$$

Thus the infinity of natural number pairs m, n is behind the definition of equality of length ratios, but only potentially so, because equality

1. It may seem unwieldy to work with ratios of lengths rather than just lengths, but in fact length is a *relative* concept and only the ratio of lengths is absolute. When we say length $a=3$, for example, we really mean that 3 is the ratio of a to the unit length. In chapter 9 we will see that the relative concept of length is a specific characteristic of Euclidean geometry.

depends on a single (though arbitrary) pair m, n. In defining unequal length ratios, infinity can be avoided completely, because one *particular* pair can witness inequality. Namely, if $a/b < c/d$ then there is a particular m/n such that

$$\frac{a}{b} < \frac{m}{n} < \frac{c}{d},$$

and likewise, if $c/d < a/b$ then there is a particular m/n between c/d and a/b. Today we would say that ratios of lengths are *separable* by ratios of natural numbers.

The Archimedean Axiom

The assumption that natural number ratios separate ratios of lengths is equivalent to a property later called the *Archimedean property*: if $a/b > 0$ then $a/b > m/n > 0$ for some natural numbers m and n. It follows, obviously, that in fact $a/b > 1/n$, so $na > b$. This gives the usual statement of the **Archimedean axiom**: *if a and b are any nonzero lengths, then there is a natural number n such that* $na > b$.

Another statement of the Archimedean axiom is: *there is no ratio* a/b *so small that* $0 < a/b < 1/n$ *for each natural number n*, or more concisely, *there are no infinitesimals*. This property was assumed by Euclid and Archimedes (hence the name), but some later mathematicians, such as Leibniz, thought that infinitesimals exist. We will see in chapter 4 that the existence of infinitesimals was a big issue in the development of calculus.

Mathematical practice today has translated Eudoxus's theory into our concept of the **real number system** $\mathbb{R}$. The ratios of lengths are the nonnegative real numbers, and among them lie the nonnegative **rational numbers**, which are the ratios m/n of natural numbers. Any two distinct real numbers are separated by a rational number, so there are no infinitesimals in $\mathbb{R}$. Conversely, each real number is determined by the rational numbers less than it and the rational numbers greater than it. Exactly how this came about, and what the real numbers *are*, is explained in chapter 11. It turns out that separation by rational numbers is the key to answering this question.

The Method of Exhaustion

We discuss the method of exhaustion only briefly here, because it is a generalization of the theory of proportions. Also, the best examples of the method occur in the work of Euclid and Archimedes, discussed in chapter 2. The basic idea is to approximate an "unknown quantity," such as the area or volume of a curved region, by "known quantities" such as areas of triangles or volumes of prisms. This generalizes the idea of approximating a ratio of lengths by ratios of natural numbers. Generally, there is a potential infinity of approximating objects, but as long as they come "arbitrarily close" to the unknown quantity it is possible to draw conclusions without appealing to actual infinity.

An example is approximation of the circle by polygons, shown in figure 1.9, which allows us to draw the conclusion that the area of the circle is proportional to the square of its radius.

Figure 1.9 shows polygons approximating the circle from inside and outside. Only the first two approximations are shown, but one can imagine a continuation of the sequence by repeatedly doubling the number of sides. It is clear that the area of the gap between inner and outer polygons becomes arbitrarily small in the process, and hence both inner and outer polygons come arbitrarily close to the circle in area.

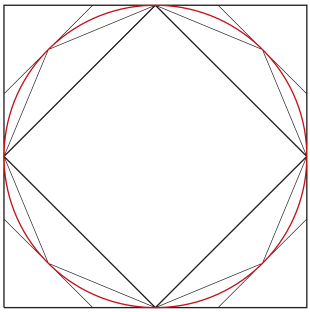

Figure 1.9 : Approximating the circle by polygons

Also, the area of each polygon P_n is a sum of triangles, whose area $P_n(R)$ for radius R is known and proportional to R^2. Now comes a typical example of reasoning "by exhaustion": suppose that the area $C(R)$ of the circle of radius R is *not* proportional to R^2. Thus, if we compare circles of radius R and R' we have either

$$C(R)/C(R') < R^2/R'^2$$

or

$$C(R)/C(R') > R^2/R'^2.$$

If $C(R)/C(R') < R^2/R'^2$, then by choosing n so that $P_n(R)$ is sufficiently close to $C(R)$ and $P_n(R')$ is sufficiently close to $C(R')$, we will get

$$P_n(R)/P_n(R') < R^2/R'^2,$$

which is a contradiction. If $C(R)/C(R') < R^2/R'^2$ we get a similar contradiction. Therefore *the only possibility is that* $C(R)/C(R') = R^2/R'^2$.

We have established what we want by *exhausting* all other possibilities. This is what "exhaustion" means in the method of exhaustion. Notice also that we used only the potential infinity of polygons by going only far enough to contradict a given inequality. This is typical of the method.

1.7 REMARKS

We have seen in the development of Greek mathematics many topics considered tricky in undergraduate mathematics today, such as proof by contradiction, the use of infinity, and the idea of choosing a "sufficiently close" approximation. This just goes to show, in my opinion, that ancient mathematics is good training in the art of proof.

At the same time, we have seen that ancient arguments can often be streamlined by the use of algebraic symbolism, and the art of algebra was missing in ancient times.

The other thing missing, in what we know of this early stage, was the systematic deduction of theorems from axioms. The art of **axiomatics** also began in ancient times, as we will see in the next chapter.

Index