# [TOPICS IN PRE-CALCULUS]
# Functions, Graphs, and Basic Mensuration Formulas

H. Wu

January 11, 1995; **revised June 15, 2003**

The first purpose of this article is to give a careful review of the concept of a function and its graph, and to give a **proof** of the fact that the graph of a linear function is a straight line. This fact is traditionally used in school mathematics without any explanation, but its proof holds the key to the understanding of the many forms of the equation of a straight line. A second purpose is to give careful proofs of all the standard area formulas of rectilinear plane figures, the point here being that the explanations of these formulas in most textbooks are incomplete. In the process, the meaning of the number $\pi$ and the area formula of a circle will also be clarified. The article concludes with some general comments about volume, particularly the relationship between the volumes of generalized cylinders and cones which have the same base and the same height.

Basic to calculus (and all of mathematics and science) is the concept of a *function*. We first recall its definition.

Given two sets $D$ and $R$, let us first consider some rules (or procedures) that assign to each element of $D$ an element of $R$.

> *Example 1.* Let $D$ be the set of all human beings who have at least one sister, and $R$ be the set of all human beings. let $F_1$ be the rule that assigns to each person $\Box$ in $D$, $\Box$'s sister. Symbolically:
>
> $$F_1 : \Box \longrightarrow \Box\text{'s sister}$$
>
> If $\Box$ has several sisters, the assignment specified by this rule is then ambiguous: we could assign to $\Box$ any of the sisters, and it would be unclear which one of $\Box$'s sisters $F_1$ assigns.

*Example 2.* Again let $D = R =$ the set of all human beings, and let $F_2$ be the rule that assigns to each $x$ in $D$, $x$'s biological father. This assignment

$$F_2 : x \longrightarrow x\text{'s biological father}$$

is then meaningful and unambiguous, except in the case where $x$'s biological father is no longer living. Thus if $x$'s father is no longer alive, $F_2$ cannot assign anyone in $R$ to $x$. Now we have the situation that there are elements (people) in $D$ to which $F_2$ cannot assign any element (person) in $R$.

*Example 3.* Let now $D$ be the set of all human beings whose biological fathers are still living, and let $R$ be the set of all human beings. Now let $F_3$ be the rule that assigns to each $\clubsuit$ in $D$ $\clubsuit$'s biological father. This rule $F_3$ then makes sense for each and every $\clubsuit$ in $D$ and furthermore, this rule is totally unambiguous: $F_3$ assigns to each $\clubsuit$ in $D$ one and only one element of $R$. We express this by saying that $F_3$ *is well-defined on* (all of) $D$.

*Example 4.* Let $D$ and $R$ be the real numbers and let $F_4$ be the rule that assigns to each real number $\square$ a square root of the sum of 1 and the square of $\square$. Thus,

$$F_4 : \square \longrightarrow \pm\sqrt{1 + \square^2}.$$

Then the rule $F_4$ is ambiguous because, for example, it could assign to 1 either $+\sqrt{2}$ or $-\sqrt{2}$.

*Example 5.* Let $D$ and $R$ be the real numbers and let $F_5$ be the rule that assigns to each real number $\square$ the negative square root of the sum of 1 and the square of $\square$. Thus,

$$F_5 : \square \longrightarrow -\sqrt{1 + \square^2}.$$

The rule $F_5$ is then totally unambiguous for every $\square$ in $D$, and we again say that $F_5$ is well-defined on $D$.

*Example 6.* Let $D$ and $R$ be the real numbers and let $F_6$ be the rule that assigns to each real number its reciprocal. Thus

$$F_6 : \square \longrightarrow 1/\square.$$

However, $F_6$ cannot assign a real number in $R$ to 0 because 1/0 has no meaning. (Recall, $\infty$ is not a real number.) But if we change $D$ to $D_6 \equiv$ {the set of all nonzero real numbers}, then $F_6$ as a rule from $D_6$ to $R$ is now well-defined on all of $D_6$.

We thus see from the above examples that some rules may seem to do a good job of assigning to each element of $D$ an element of $R$, but fail to do so upon closer scrutiny because either the assignment cannot be made on certain elements of $D$ or the assignment is ambiguous because it assigns to an element of $D$ one of *several* possible elements in $R$. We are interested in those unambiguous rules that are meaningful on all of $D$, and these are called "funtions". Formally, a *function $F$ from a set $D$ to a set $R$* is a rule that assigns to each element $\square$ of $D$ one and only one element of $R$. Thus, a function from $D$ to $R$ is an unambiguous assignment that can be made on all of $D$. For obvious reasons, *the* element assigned to $\square$ by $F$ is denoted by $F(\square)$. In symbols: it is convenient to denote the function by

$$F : D \to R,$$

and we also write

$$F : \square \longrightarrow F(\square).$$

In the preceding examples, $F_3$ and $F_5$ are functions, and so is $F_6$ when considered as a rule from $D_6$ to $R$. The rules in the remaining examples are not functions.

Because the concept of a function has been around for so long, every aspect of the above definition has acquired a name of its own and you will have to get used to the terminology. So with $F : D \to R$ given:

$D$ is the *domain* of $F$

$R$ is the *range* of $F$

the elements $\square$ in $D$ are the *variables*

$F(\square)$ is the *value* of $F$ at $\square$

$F$ is *defined on* $D$

$F$ *takes values* in $R$

For our purpose, the following two kinds of functions are the most important. If $D$ and $R$ are subsets of the *reals* (= real numbers), then a function $F : D \to R$ is called a *real-valued function of a real variable*. If $D$ is a subset of the plane (consisting of all ordered pairs $\{(x, y)\}$ of real numbers), and $R$ is a subset of the reals, then $F : D \to R$ is called a *real-valued function of two variables* $x$ *and* $y$. Since we almost never consider any function which does not take value in the reals, we simply refer to them as *a function of one variable* and *a function of two variables*, respectively.

CONVENTION: *If a real-valued function of a real variable $F$ is given, it will always be assumed that its domain is the set of all real numbers on which $F$ is well-defined.*

For example, the domain of the function $F : \square \longrightarrow 1/\square$ of one variable, without further specification, would be understood to have the set of all nonzero numbers as its domain (cf. Example 6 above). Similarly, the domain of the function $H : x \longrightarrow +\sqrt{1-x}$, without further comments, would be assumed to be the semi-infinite closed interval $(-\infty, 1]$ (i.e., all the real numbers $x$ so that $x \leq 1$).

Until further notice, we shall consider only functions of one variable. Recall the definition of the *graph* of such a function $F$: it is the set of all the points $(x, F(x))$ in the plane, where $x$ belongs to the domain of $F$. The emphasis here is on the word "all". In greater detail, this means: if $C$ is the graph of $F$, then

(i) for every $x$ in the domain of $F$, the pair $(x, F(x))$ belongs to $C$, and

(ii) if a point $(x, t)$ of the plane belongs to $C$, then necessarily $x$ belongs to the domain of $F$ and $t = F(x)$.

Because $F(x)$ can be one and only one element of the range of $F$ (see the above definition of a function), we also have:

(iii) if $(a, b)$ and $(a, c)$ are two points in the graph $C$ of $F$, then $b = c$ (both being equal to $F(a)$).

The last property (iii) leads to the following *vertical line test* for the graph of a function. (Recall: a line is *vertical* if it is parallel to the $y$-axis. This is the same as saying that there is a constant $a$, so that all the points on the line have coordinates $(a, y)$, where $y$ is a real number.)

**Theorem 1**. A nonempty subset $C$ of the plane is the graph of a function if and only if every vertical line intersects $C$ in at most one point.

Before proving this fact, let us first explain the meaning of the phrase "*if and only if* ". This is a common mathematical shorthand that combines two statements into one and it is very important to have a thorough understanding of this phrase. Given statements **A** and **B**, the sentence "**A** if and only if **B**" means precisely that both of the following implications are true:

(1) **A** implies **B**, and

(2) **B** implies **A**.

Thus to prove Theorem 1, we need to prove two statements:

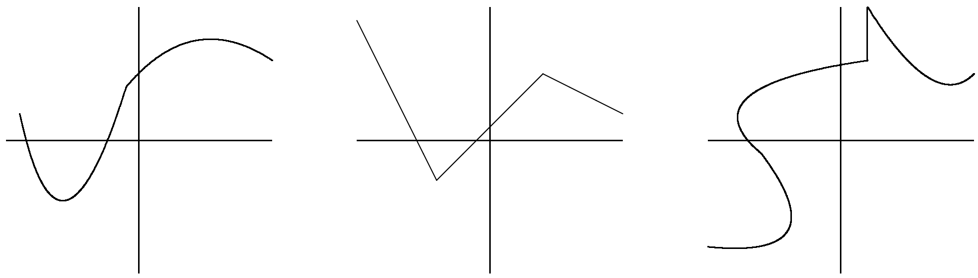
(a) if $C$ is the graph of a function, then every vertical line intersects $C$ in at most one point, and
(b) if every vertical line intersects $C$ in at most one point, then $C$ is the graph of a function.

We first prove (a). Let $C$ be the graph of a function $F : D \to R$ and let $L$ be a vertical line passing through a point $\Box$ on the $x$-axis. Thus all points on $L$ have coordinates $(\Box, t)$, where $t$ is an arbitrary number. If $\Box$ does not belong to $D$, then no point of $L$ could lie in $C$, by (ii) above. Then $L$ intersects $C$ in no (zero) point. If $\Box$ belongs to $D$, then $L$ intersects $C$ in exactly one point, namely, $(\Box, F(\Box))$, by (i) and (iii). This proves (a). To prove (b), denote by $L_t$ the vertical line passing through the point $(t, 0)$ on the $x$-axis. Let $D$ be the set of all the $t$'s so that $L_t$ intersects $C$ in exactly one point. This means that if $t \in D$, then the vetical line $L_t$ passing through the point $(t, 0)$ on the $x$-axis meets $C$ at a point $(t, T)$ for some real number $T$, and that for an $s$ not in $D$, The vertical line $L_s$ passing through the point $(s, 0)$ on the $x$-axis would not meet $C$. Using this notation, we can now define the function $F$ from $D$ to the reals by: for any $t \in D$,

$$F : t \longrightarrow \text{the } y\text{-coordinate } T \text{ of the point of intersection } (t, T) \text{ of } L_t \text{ with } C.$$

Because there is *exactly one* $T$ so that $(t, T) \in C$, this rule defines a function $F$ whose graph lies in $C$. Moreover, if $(t', T') \in C$, then the vertical line passing through $(t', 0)$ on the $x$-axis intersects $C$ at $(t', T')$ and, by hypothesis, at no other point of $C$. Therefore, $(t', T') = (t', F(t'))$, and therefore, every point of $C$ is in the graph of $F$. It follows that $C$ is exactly the graph of $F$. Q.E.D. (="End of proof")

As illustrations of Theorem 1, we can tell right away that the first two curves below are graphs of functions, whereas the third one is not:



Next, we look at some common examples of functions.

Functions of the form

$$F : \Box \longrightarrow a\Box + b,$$

where $a$ and $b$ are real numbers, occupy a position of singular importance. These are called *linear functions*. We proceed to review some of the basic facts and relevant terminology.

The following theorem is central to the understanding of the relationship between a linear function and its graph.

**Theorem 2**. The graph of a linear function is a straight line.

This fact accounts for the nomenclature of a *linear* function.

Theorem 2 is seldom proved or even explicitly mentioned in elementary texts. Let us first make clear what is involved. Take two distinct points $A$ and $B$ on the graph $G$ of a given linear function $F : \Box \longrightarrow a\Box + b$, and let $L$ be the straight line joining $A$ and $B$. If Theorem 2 is true, then obviously $G$ must coincide with $L$. (Here we are using, in a crucial way, the fact that through two distinct points passes one and only one straight line.) Equally obviously, if we can show that $G = L$, then $G$ is a straight line, namely, $L$. Therefore, our strategy is to make a judicious choice of two points $A$ and $B$ on $G$ and prove that the line $L$ so defined satisfies $L = G$. Now how does one prove that two sets $L$ and $G$ are equal? It is time to recall that, *by definition*, the equality of the two sets, $L = G$, means exactly that the following two statements are true simultaneously:

($\alpha$) every element in $L$ is an element of $G$, and

($\beta$) every element in $G$ is an element in $L$.

*Thus proving $L = G$ is equivalent to proving these two statements ($\alpha$) and ($\beta$).* (Keep this in mind and it will save you a lot of grief later on in your other math courses.)

We begin with an *observation*: Let $A$ and $B$ be two arbitrary points in the plane with distinct $x$-coordinates. Let $L$ be the line joining $A$ and $B$. Then $L$ is not vertical. Let the coordinates of $A$ and $B$ be $(x_1, y_1)$ and $(x_2, y_2)$, respectively; it is then customary to write the points as $A(x_1, y_1)$ and $B(x_2, y_2)$ to indicate their coordinates. Because $x_1 \neq x_2$, we may introduce the number $m$ defined by

$$m \equiv \frac{y_2 - y_1}{x_2 - x_1}. \tag{1}$$

Notice that $m$ is the quotient of the difference of the $y$-coordinates of $A$ and $B$ divided by the difference of the $x$-coordinates of the same, *in the order indicated*, $B$ before $A$. But it is easy to see that we could equally well let $A$ precede $B$ in the definition because

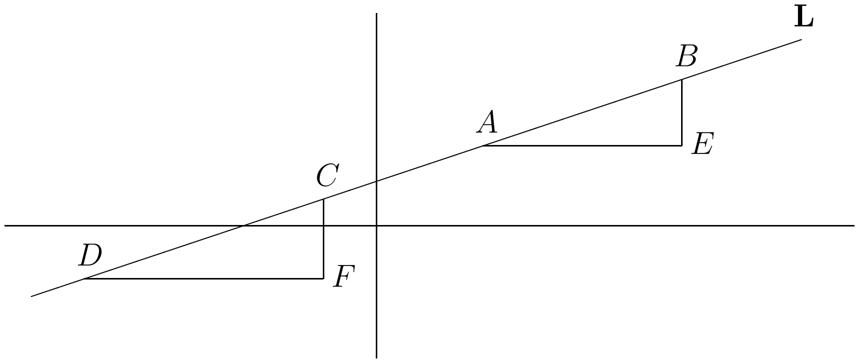
$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_1 - y_2}{x_1 - x_2}.$$

We wish to prove that this $m$ is in fact a number attached to the line $L$ itself and not just to the points $A$ and $B$. In other words, if we take *any* two points $C(X, Y)$ and $D(X', Y')$ on $L$, then we would also get

$$m = \frac{Y' - Y}{X' - X} \tag{2}$$

We call this $m$ the *slope* of $L$. The fact that the slope of a line $L$ can be computed using *any* two points on the line is usually not made explicit, much less explained, in the school curriculum, thereby creating an unnecessary obstacle to the learning of algebra.

The proof of equation (2) depends on the theory of similar triangles. We refer to the picture with $L$ as shown. The proof would proceed exactly the same way if $L$ is slanted the opposite way to the left, i.e., $\setminus$ .




Form triangles $BAE$ and $CDF$ so that $BE$ and $CF$ are parallel to the $y$-axis while $EA$ and $FD$ are parallel to the $x$-axis. Then from the equality of the respective angles, we have the similar triangles $\triangle BEA \sim \triangle CDF$. If $|BE|$ denotes the length of the line segment $BE$, etc., we have $|BE|/|EA| = |CF|/|FD|$. In terms of the respective coordinates, we see from the picture that this is the same as

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{Y - Y'}{X' - X},$$

which is the same as

$$\frac{Y' - Y}{X' - X} = \frac{y_2 - y_1}{x_2 - x_1} = m,$$

which is equation (2). (This argument appears to depend on the particular ordering of the points $B$, $A$, $C$ and $D$ on $L$ as shown in the picture. The fact that were their ordering different the same argument would have led to the same conclusion will be left as a simple exercise to the reader.)

We can now prove Theorem 2. We are given a linear function $F$ so that $F(x) = ax + b$, where $a$ and $b$ are constants. We have to show that the graph $G$ of $F$ is a straight line. Now if $a = 0$, then $F(x) = b$ for all $x$, and the graph $G$ would consist of all the points of the form $(x, b)$ in the plane, where $x$ is arbitrary. In this case, $G$ is obviously the "horizontal" line, parallel to the $x$-axis, passing through $(0, b)$. We may threrefore assume $a \neq 0$.

Observe that $A(0, b)$ and $B(-\frac{b}{a}, 0)$ are two points belonging to $G$ (in fact, $b$ and $-\frac{b}{a}$ are the so-called *y-intercept* and *x-intercept* of $G$, respectively). Let $L$ be the line joining $A$ and $B$. We shall prove $L = G$. This means we have to prove ($\alpha$) and ($\beta$) above. We first prove ($\alpha$). Thus we must show that every point of $L$ belongs to $G$. Let $C(X, Y) \in L$. To show that $C(X, Y) \in G$, we have to show that $Y = F(X)$, i.e.,

$$Y = aX + b \tag{3}$$

By the observation concerning the slope of a line, we may compute the slope of $L$ using either $A$ and $C$, or $A$ and $B$, and the two numbers must be equal. Thus

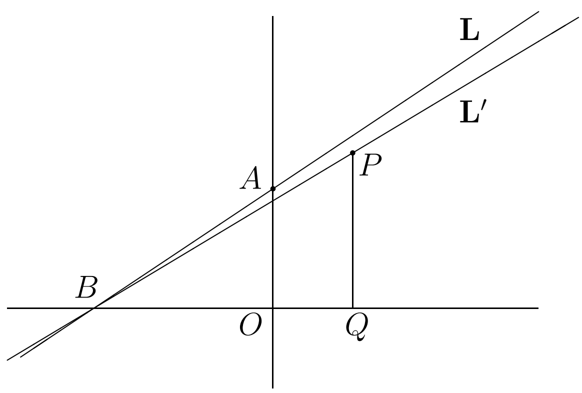
$$\frac{b - Y}{0 - X} = \frac{b - 0}{0 - (-\frac{b}{a})},$$

i.e.,

$$\frac{Y - b}{X} = a,$$

which is exactly (3). So ($\alpha$) is true. (Incidentally, we have proved that the slope of a line passing through the points $A(0, b)$ and $B(-\frac{b}{a}, 0)$ is $a$.)

Next we prove ($\beta$), i.e., if $P \in G$, then $P \in L$. Since $G$ is the graph of the function $F : x \longrightarrow ax + b$, we may let $P = P(X, aX + b)$ for some number $X$. Suppose $aX + b > 0$. We let $L'$ be the straight line joining $P$ to $B$, as shown:




If we can show $L = L'$, then it would follow that $P \in L$. Let $Q$ be the point on the $x$-axis so that $PQ$ is parallel to the $y$-axis, and let $O$ be the origin of the coordinate system. Consider the triangles $PBQ$ and $AOB$. Let $|PQ|$ denote the length of the line segment $PQ$, etc. Then

$$\frac{|PQ|}{|QB|} = \frac{aX + b}{X + \frac{b}{a}} = a$$

and

$$\frac{|AO|}{|OB|} = \frac{b}{\frac{b}{a}} = a.$$

Therefore

$$\frac{|PQ|}{|QB|} = \frac{|AO|}{|OB|},$$

and since angles $\angle AOB$ and $\angle PQB$ are right angles, the SAS criterion for similar triangles implies $\triangle ABO \sim \triangle PBQ$. Therefore $\angle ABO$ is equal to $\angle PBQ$. But these two angles share a side (the half line from $B$ to $Q$) and they are both in the upper half plane, so their other sides must coincide, i.e., the lines $AB$ and $PB$ coincide. Thus we have proved $L = L'$, and therefore $P \in L$ in case $P$ lies in the upper half plane (i.e., $aX + b > 0$). The case of $P$ being in the lower half plane (i.e., $aX + b < 0$) is entirely similar. Q.E.D.

It is clear that the truth of Theorem 2 depends crucially on the theory of similar triangles. This fact should be emphasized in the school curriculum. Sadly, such is currently not the case.

If we rewrite equation (1) as

$$m = \frac{y - y_1}{x - x_1}$$

with $(x_1, y_1)$ held fixed and $(x, y)$ ranging over the line $L$, it is called the *point-slope form* of the equation of the line $L$. If the two points $(x_1, y_1)$ and $(x_2, y_2)$ are given on $L$ and we want to characterize all the points $(x, y)$ on $L$, then we use the fact that the slope of $L$ can be computed by any two points to conclude that

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}.$$

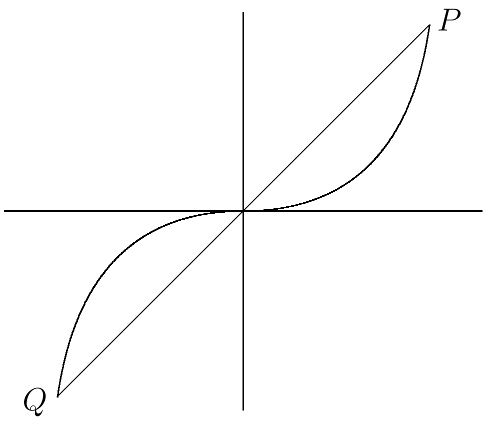
This is then called the *two-point form* of the equation of the line $L$. Finally, suppose the line $L$ meets the $y$-axis and $x$-axis at $A(0, b)$ and $B(-\frac{b}{a}, 0)$, respectively. Then for an arbitrary point $(x, y)$ on $L$, we have from equation (3) that

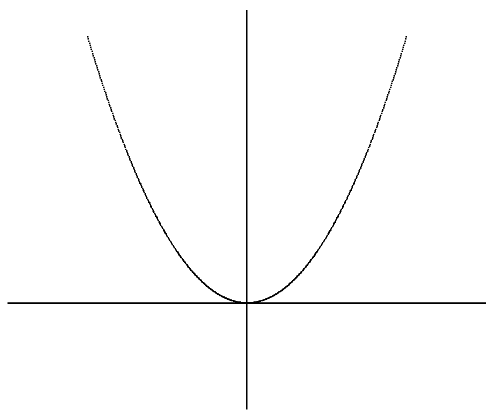
$$y = ax + b.$$

This is called the *slope-intercept form* of the equation of $L$ because we saw in the proof above that $a$ is the slope of $L$. When the proof of Theorem 2 is understood, all three forms become parts of a coherent picture. Theorem 2 therefore provides the context to put all three forms in the proper perspective. Without a knowledge of the proof of Theorem 2, these three forms become three isolated facts to be memorized by brute force.

To round off the picture, let us briefly recall some basic facts and terminology associated with a straight line. Two distinct lines are parallel if and only if they have the same slope, and are perpendicular if and only if the product of their slopes is $-1$ (in other words, one slope is the negative of the reciprocal of the other). For lack of space, we won't go into the explanation of these facts. Once you understand the preceding reasoning, however, their explanation is straightforward (but be sure you know it). It is worth noting that the sign of the slope $m$ of a line $L$ has geometric significance: if $m > 0$, then $L$ is slanted this way $/$, whereas if $m < 0$, then $L$ is slanted in the other direction $\backslash$. Moreover, the absolute value of the slope also has geometric significance: if $|m|$ is large, then $L$ is close to being vertical whereas if $|m|$ is small, then $L$ is close to being horizontal. Again, you should try to supply the easy reasoning.

We should bring to your attention two general classes of functions and their graphs. A function $F$ is *odd* if for every $x$ in the domain of $F$, $F(x) = -F(-X)$, and is *even* if for every $x$ in the domain of $F$, $F(x) = F(-x)$. If $F : x \longrightarrow x^n$, then $F$ is odd exactly when $n$ is an odd integer, and is even exactly when $n$ is an even integer. In terms of their graphs, the graph of an odd function (say defined for all real numbers) is *radially symmetric* with respect to the origin of the plane, in the sense that if a point $P$ is on the graph of an odd function $F$, then if we join $P$ to the origin and then extend this line segment to the same length on the other side of the origin, we would land on a point $Q$ which will again lie on the graph of $F$. The reason is simple. If the coordinate of $P$ are $(a, F(a))$, then the coordinates of $Q$ must be $(-a, -F(a))$ due to the fact that $P$ and $Q$ are by choice diametrically opposite with respect to the origin. Since $F$ is odd, $-F(a) = F(-a)$, so that the coordinates of $Q$ are in fact $(-a, F(-a))$, which then exhibits $Q$ as a point on the graph of $F$. If $F$ is even instead, then its graph would be symmetric with respect to the $y$-axis, for quite similar reasons. Both kinds of behavior are displayed in the following graphs of $F : x \longrightarrow x^n$ for $n$ odd and $n$ even, respectively.



It should also be pointed out in this connection that every function can be written as the sum of two functions, one odd and the other even. The way to do it is a standard trick that is worth learning. If $F$ is any function, the functions $F_1$ and $F_2$ defined respectively by:

$$F_1(x) = \frac{1}{2}(F(x) + F(-x)) \quad \text{and} \quad F_2(x) = \frac{1}{2}(F(x) - F(-x))$$

for any number $x$ is easily seen to be even and odd, respectively. Equally obvious is the fact that

$$F(x) = F_1(x) + F_2(x).$$

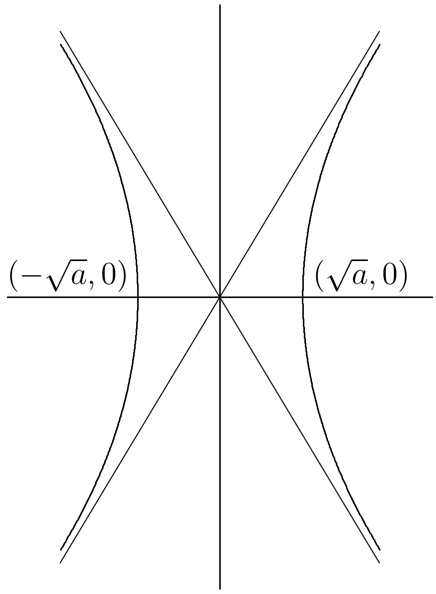
This is the desired sum.

We now discuss graphs of equations of two variables, such as $x^2 + y^2 = 1$. Let a function of two variables $f : (x, y) \longrightarrow f(x, y)$ be given. Then the *graph of the equation* $f(x, y) = 0$ is by definition the set of *all* the points $\{(X, Y)\}$ in the plane so that $f(X, Y) = 0$. In this context then, what we called "the graph of $F : \Box \longrightarrow F(\Box)$" above is nothing but the graph of the equation $g(x, y) = 0$, where $g$ is the function of two variables $g : (x, y) \longrightarrow y - F(x)$. In the same way, what we called above "the graph of $x^2 + y^2 = 1$" is nothing but the graph of the equation $h(x, y) = 0$, where $h$ is the function $h : (x, y) \longrightarrow x^2 + y^2 - 1$. Note that, by tradition, it is much more common to speak of "the graph of $x^2 + y^2 = 1$" rather than "the graph of the equation $x^2 + y^2 - 1 = 0$". For this reason, we shall henceforth follow the traditional usage in this and other similar situations.

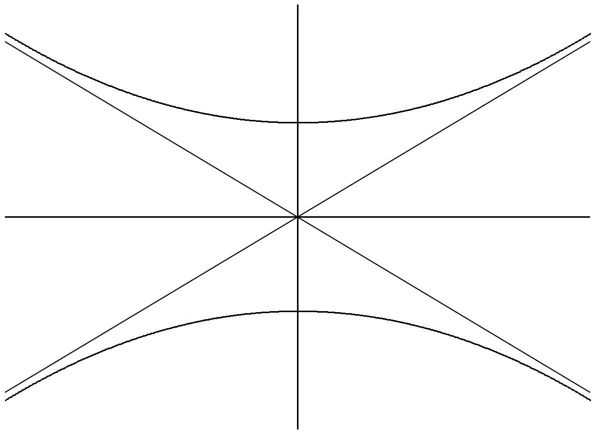
There are several standard graphs of equations that you should know about. First of all, the graph of $ax + by = c$, where $a$, $b$ and $c$ are arbitrary real numbers so that not both $a$ and $b$ are zero, is of course a straight line. Indeed, if $b = 0$, then the equation reduces to $x = c/a$, which is the vertical line consisting of all the points $\{(c/a, t)\}$, $t$ real. If $b \neq 0$, then the equation becomes $y = -(a/b)x + (c/b)$, and we recognize from equation (3) above that this is a straight line with slope $-a/b$ and $y$-intercept $c/b$.

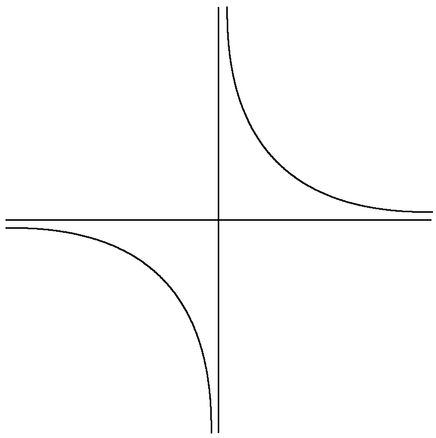
The graph of $x^2 + y^2 = r$ for any positive $r$ is of course the circle of radius $\sqrt{r}$. The graph of $\dfrac{x^2}{a} + \dfrac{y^2}{b} = 1$ where $a$ and $b$ are distinct positive numbers is an *ellipse* in the so-called *standard position* which passes through the points $(\sqrt{a}, 0)$ and $(0, \sqrt{b})$ on the $x$-axis and $y$-axis, respectively. It is a "squashed circle" if $a > b$, and is an "elongated circle" if $a < b$.

The graph of $\dfrac{x^2}{a} - \dfrac{y^2}{b} = 1$ where $a$ and $b$ are positive numbers is a pair of *hyperbolas* passing through the points $(\pm\sqrt{a}, 0)$. The lines $y = \pm\left(\sqrt{\frac{b}{a}}\right) x$ are the so-called *asymptotes* of the hyperbolas, meaning that the the hyperbolas get arbitrarily close to these straight lines when they are sufficiently far away from the origin but never intersect them.



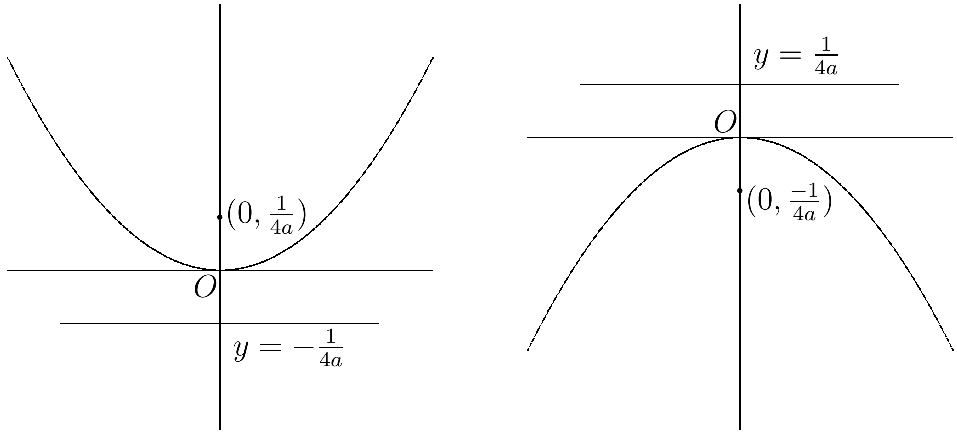

There are two variants of the above equation for hyperbolas. One is $\frac{y^2}{a} - \frac{x^2}{b} = 1$. Clearly this is merely a matter of exchanging the roles of $x$ and $y$ (see the first of the figures below.) The other is $xy = k$, where $k > 0$. These hyperbolas have the coordinate axes as asymptotes (see the second of the figures below).





Note that one can "transform" the first two kinds of hyperbolas into one like the third kind (i.e., $xy = k$) through a rotation and a "shear" of the plane, but such an explanation would require a more detailed discussion than is possible here. In any case, such a discussion can be found in all the linear algebra courses and texts.

Finally, the graph of $y = ax^2$ $(a > 0)$ is a *parabola in standard position.* It is worth mentioning that this parabola is exactly the set of all points equi-distant from the point $(0, 1/(4a))$ and the horizontal line $y = -1/(4a)$. (Discovering this fact on your own may be difficult, but once you have been told it is true, the verification is a straightforward computation that you should carry out yourself.) The former is called the *focus* of the parabola, and the latter its *directrix.* If $a < 0$, then we get the upside-down version of the former.




We now recall some area and volume formulas.

Without defining what "area" and "volume" mean (these are subtle concepts, but you will get a glimpse of the correct definitions in calculus), we shall make some simple assumptions and use them to derive some area formulas for rectilnear figures in the plane.

Let us start with the area of a square of side $a$: we would all agree that it should be $a^2$. Next,

$$\text{(the area of a rectangle of sides } a \text{ and } b) \;=\; ab.$$

Although a little less obvious than the case of the square, let us also accept this. We will also *assume* that

(i) congruent regions have the same area, and

(ii) if a region is expressed as the union of two sub-regions which intersect each other at most on their boundaries, then the area of the region is the sum of the areas of the sub-regions.

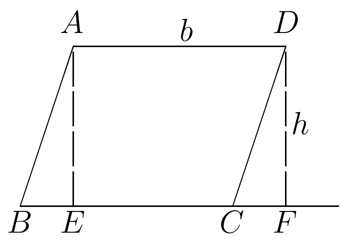
These are completely believable statements and are usually taken for granted.

We now prove by drawing pictures that

$$\text{(the area of a parallelogram with base } b \text{ and height } h) \;=\; bh. \tag{4}$$

In equation (4), it is understood that the height is taken with respect to the given base.

Indeed, let $ABCD$ be the given parallelogram and let $BC$ be the given base with length $b$. Note that the length of the side $AD$ is then also equal to $b$. Let prependiculars $AE$ and $DF$ be dropped on $BC$. There are two possibilities: whether at least one of $E$ and $F$ falls on the segment $BC$ or both fall outside the segment $BC$. We first assume the former, as shown:

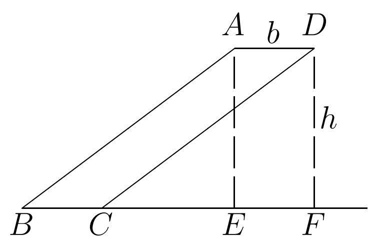



Then the triangles $ABE$ and $DCF$ are clearly congruent so that they must have the same area (using (i)). Moreover $AEFD$ is a rectangle and therefore has area $bh$. Thus,

$$\begin{aligned} \text{area of } ABCD &= \text{area of } ABE + \text{area of } AECD \\ &= \text{area of } DCF + \text{area of } AECD \\ &= \text{area of } AEFD \quad = \quad bh, \end{aligned}$$

as claimed.

We now tackle the second case where both $E$ and $F$ fall outside the base $BC$, as shown:




The triangles $ABE$ and $DCF$ are congruent (by ASA), and therefore have the same area. Moreover $AEFD$ is a rectangle and its area is $bh$. Thus,

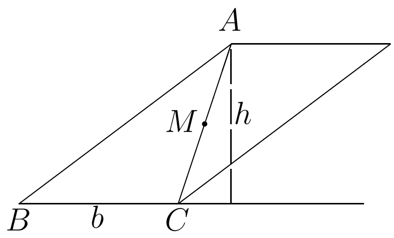
$$\begin{aligned} \text{area of } ABCD &= \text{area of } ABFD - \text{area of } DCF \\ &= \text{area of } ABFD - \text{area of } ABE \\ &= \text{area of } AEFD \quad = quadbh, \end{aligned}$$

So equation (4) has now been completely proved. (Note how assumptions (i) and (ii) above were used implicitly but repeatedly in the proofs of both cases.)

Using equation (4), we can now prove that

$$\text{the area of a triangle with base } b \text{ and height } h \quad = \quad \frac{1}{2}bh. \tag{5}$$

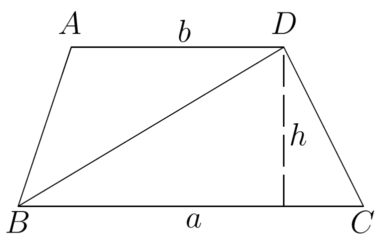
This is also proved by a picture (see below): Given triangle $ABC$, let $M$ be the midpoint of segment $AC$. Rotate the triangle 180 degrees around $M$ to get an inverted copy of $\triangle ABC$ adjoining side $AC$. Together, these two triangles now form a parallelogram. The latter has twice the area of the triangle, but according to equation (4), it also has area $bh$. So by assumptions (i) and (ii) again, (5) follows.




Equation (5) has important applications. On the one hand, it can be used to derive the area of a trapezoid (recall: a *trapezoid* is a quadrilateral with a pair of parallel opposite sides). Let the lengths of the parellel sides be $a$ and $b$, and let the height (= the distance between the parallel sides) be $h$. Then,

$$\text{area of trapezoid} \quad = \quad \frac{1}{2}h(a+b). \tag{6}$$

Let the trapezoid be $ABCD$ (see picture below). The diagonal $BD$ divides it into two triangles, $\triangle ABD$ and $\triangle BCD$. By assumption (ii), the area of $ABCD$ is the sum of the areas of $\triangle ABD$ and $\triangle BCD$. Now apply equation (5) to both triangles and we immediately get (6).




A more important application of equation (5) is of a theoretical nature. Every polygon is the union of triangles which intersect each other at most at boundary edges; although quite difficult to prove in general, this fact is altogether believable and will therefore be taken for granted. In the case most important to us, that of a regular polygon, it is obvious how to express it as such a union of triangles. In any case, because the area of a triangle can be computed by (5), the area of *any* polygon is therefore computable in principle. We will see why this remark is important when we come to the area of disks below.

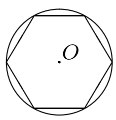
We now come to something much more delicate: the area of the region inside a circle. First, a word about terminology. In every day conversation and, sadly, also in most high school texts as well as in most calculus texts, this would be called "the area of a circle", but the resulting confusion (does "circle" refer to the curve or the region inside this curve?) is clearly unaccceptable. So in mathematics at least, we make a distinction: henceforth *circle* refers only to the round curve, and *disk* will designate the region inside the circle. Thus a circle of radius $r$ is the boundary of a disk of radius $r$. This being said, we are going to give a correct definition of the number $\pi$, and then give an intuitive explanation of the formula:

$$\text{the area of a disk of radius } r \quad = \quad \pi r^2. \tag{7}$$

More importantly, we also give an intuitive explanation of

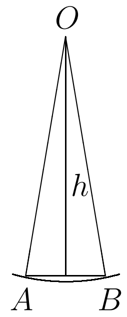
$$\text{the length of a circle of radius } r \quad = \quad 2\pi r \tag{8}$$

The number $\pi$ is by definition *the area of the disc of radius* 1. We now give a restrictive, but correct, definition of the area of a disk of any radius $r$, to be denoted by $D(r)$. Let the boundary circle of $D(r)$ be denoted by $C(r)$, and let $P_n$ be a regular $n$*-gon* (a polygon of $n$ sides) inscribed in $C(r)$, i.e., all the vertices of $P_n$ lie on $C(r)$. The following shows the case of $n = 6$ for a disk with center $O$:




It is intuitively clear that as $n$ gets large without bound, $P_n$ would get closer and closer to $C(r)$ and the region inside $P_n$ would become virtually indistinguishable from the disk $D(r)$. Because the area of $P_n$ is something we know how to compute, it makes sense then to *define the area of* $D(r)$ *to be the* limit *of the area of* (*the region inside*) $P_n$ *as* $n$ *increases to infinity.* Here we leave vague the precise meaning of "limit", except to point out that it is one of the central concepts of advanced mathematics.

Next we explain why $C(1)$, the circle of radius 1, has length $2\pi$. This statement reminds us that we have not yet given meaning to "length of $C(r)$". Using the notation of the preceding paragraph, we *define the length of* $C(r)$ *to be the limit of the perimeter of* $P_n$ *as* $n$ *increases to infinity.* We can now compute the length of $C(r)$. Letting as usual the center of $D(r)$ be $O$, let us join $O$ to two consecutive vertices $A$ and $B$ of $P_n$ to form a triangle $OAB$:



From $O$ we drop a perpendicular to side $AB$ of $\triangle OAB$ and we denote the length of this perpendicular by $h$. If we denote the length of the segment $AB$ by $s_n$ (the subscript $n$ here indicates that it is the length of one side of $P_n$), then the area of $\triangle OAB$ is $\frac{1}{2}s_n h$, by equation (5). Now remember that $\triangle OAB$ is one of $n$ congruent triangles which "pave" $P_n$, so the area of $P_n$ is $n(\frac{1}{2}s_n h)$, by virtue of assumptions (i) and (ii). We rewrite it as

$$\text{area of } P_n \;=\; \frac{1}{2}(ns_n)h$$

But $ns_n$ is the perimeter of $P_n$ because the boundary of $P_n$ consists of $n$ sides, each being congruent to the segment $AB$. So as $n$ increases to infinity, $ns_n$ becomes in the limit the length of $C(r)$, by the definition just given. Moreover, as $n$ increases to infinity, $AB$ gets smaller and smaller and therefore closer and closer to the circle $C(r)$. So the perpendicular from $O$ to $AB$ becomes $OA$ in the limit and therefore $h$ becomes the radius of $C(r)$, which is $r$. Needless to say, the area of $P_n$ becomes the area of $D(r)$, by the above definition. Letting $n$ increase to infinity, the preceding equation then becomes

$$\text{area of } D(r) \;=\; \frac{1}{2}\,(\text{length of } C(r))\, r,$$

or, what is the same,

$$\text{length of } C(r) \;=\; \frac{2}{r}\,(\text{area of } D(r)).$$

If $r = 1$, then since the area of $D(1)$ is by definition the number $\pi$, we have

$$\text{length of } C(1) \;=\; 2\pi \tag{9}$$

We can now show why (7) and (8) must be true. The key idea is to employ the concept of dilation in the plane: $\mathcal{D}_s : (x, y) \longrightarrow (sx, sy)$, where $s > 0$. The overriding fact is that $\mathcal{D}_s$ sends a curve $K$ of length $\ell$ to a curve $\mathcal{D}_s(K)$ of length $s\ell$, and a region $\mathcal{R}$ of area $A$ to a region $\mathcal{D}_s(\mathcal{R})$ of area $s^2 A$. In our present context of a circle and a disk, both can be easily explained. Since all circles of the same radius in the plane are congruent to each other — and recall that congruent figures have the same length and area — we shall henceforth assume that our circles are centered at the origin $O$. Let us start with length. $\mathcal{D}_s$ sends the circle $C(r)$ of radius $r$ to the circle $C(sr)$ of radius $sr$. If $P_n$ is an inscribing regular $n$-gon in $C(r)$, then $\mathcal{D}_s(P_n)$ is also an inscribed regular $n$-gon in $C(sr)$. The perimeter of $\mathcal{D}_s(P_n)$ is $s$ times that of $P_n$ for the following reason.

We may assume one of the triangles obtained by joining $O$ to consecutive vertices of $P_n$ is positioned as in $\triangle OAB$ of the preceding discussion, namely, the side $AB$ is parallel to the $x$-axis. Then under the dilation $\mathcal{D}_s$, $AB$ goes into a segment (parallel to the $x$-axis) of length $s$ times the length of $AB$. Since the perimeter of $P_n$ is $n$ times the length of $AB$, and likewise the perimeter of $\mathcal{D}_s(P_n)$ is $n$ times that of $\mathcal{D}_s(AB)$, the perimeter of $\mathcal{D}_s(P_n)$ is $s$ times that of $P_n$. Because the length of $C(sr)$ is by definition the limit of the perimeter of $\mathcal{D}_s(P_n)$ — and the length of $C(r)$ is likewise the limit of the perimeter of $P_n$ — it follows that the length of $C(sr)$ is $s$ times the length of $C(r)$.

Next area. Note the obvious fact that $\mathcal{D}_s(D(r))$ is $D(sr)$. So we have to prove that the area of $D(sr)$ is $s^2$ times the area of $D(r)$. Using the preceding notational setup, we observe that the area of $\mathcal{D}_s(P_n)$ is $s^2$ times the area of $P_n$, for the following reason. Because the area of $P_n$ is $n$ times the area of $\triangle OAB$, — and the area of $\mathcal{D}_s(P_n)$ is $n$ times that of $\mathcal{D}_s(\triangle OAB)$, — it suffices to show that the area of $\mathcal{D}_s(\triangle OAB)$ is $s^2$ times that of $\triangle OAB$. But the latter is obvious from formula (5): $\mathcal{D}_s$ changes each of $h$ and the length of $AB$ by a factor of $s$, and hence changes $\frac{1}{2}s_n h$ to $(\frac{1}{2}(s\, s_n)(sh)) = s^2(\frac{1}{2}s_n h)$. So we know that the area of $\mathcal{D}_s(P_n)$ is $s^2$ times the area of $P_n$. But the area of $D(r)$ is the limit of the area of $P_n$, while the area of $\mathcal{D}_s(D(r))$ is also the limit of the area of $\mathcal{D}_s(P_n)$. So the relationship

$$\text{area of } \mathcal{D}_s(P_n) \;=\; s^2\,(\text{area of } P_n)$$

implies the corresponding relationship

$$\text{area of } \mathcal{D}_s(D(r)) \;=\; s^2\,(\text{area of } D(r)),$$

which is the same as

$$\text{area of } D(sr) \;=\; s^2\,(\text{area of } D(r)),$$

which is what we want.

Now the area of the disk of radius $r$, $D(r)$, is $r^2$ times the area of $D(1)$, which is by definition equal to $\pi$. Thus the area of $D(r)$ is $\pi r^2$. This confirms (7). On the other hand, the length of the circle of radius $r$, $C(r)$, is $r$ times the length of $C(1)$, which by (9) is $2\pi$. So the length of $C(r)$ is $2\pi r$, and (8) is also true.

Finally, we have to address an issue related to our definition of $\pi$. We all know $\pi$ is something like 3.1416, and you may be puzzled by the definition of $\pi$ as the area of the disk $D(1)$ of radius 1: how to go from this definition to 3.1416? One way to estimate the area of $D(1)$ is by drawing a circle of radius 1 on a graph paper, and compare the number of grids inside the circle with that inside the unit square. If you use very fine grids, and be patient with the counting of grids, then it should be straightforward to arrive at an estimate the area of $D(1)$ to be between 3.1 and 3.2. Sometimes one can do far better. This may be the best method for getting some intuitive feeling of what $\pi$ is about. Of course the definition of $\pi$ as the area of $D(1)$, in the context of calculus, is an integral, and various standard methods can be employed to approximate the integral to as many decimal places as we wish. Even this does not give the true picture about $\pi$, however. The fact is that $\pi$ is a number that lives in almost every part of mathematics, but because it first shows up in the context of length of circles and area of disks, the common misconception is that this is the *only* way $\pi$ shows up. What is true is that there are approaches to $\pi$ which are not related to geometry and which produce different infinite series to represent $\pi$. These series yield an accurate value of $\pi$ to many decimal digits by summing only a few terms. See, for example, L. Berggren, J. Borwein and P. Borwein, **Pi: A Source Book**, Springer-Verlag, 1997. For your curiosity, here is $\pi$ to 50 decimal places:

$$3.14159265358979323846264338327950288419716939937510$$

At least one other area formula in connection with $\pi$ must be mentioned:

$$\text{the area of a sphere of radius } r \;=\; 4\pi r^2 \qquad (10)$$

As in the case of the circle, here we mean by *sphere* the 2-dimensional round surface; we shall use *ball* to designate the 3-dimensional region inside a sphere. The remarkable nature of formula (10) should escape no one: it implies that the area of the *hemisphere* (one-half of a sphere) on top of a disk of radius $r$ is exactly *twice* that of the disk (by virtue of (7)). Given a hundred chances, would you have guessed this correctly if you hadn't known formula (10) beforehand? Needless to say, (10) requires calculus for its proof, but unlike (7), there is no similar simple explanation. In fact, formula (10), which was first discovered by Archimedes (287-212 B.C.), is among the deepest theorems of elementary mathematics. (In order to appreciate the depth of formula (10), just stop to think for a second: what is really meant by the "area" of such a *curved* surface??)

Finally we turn to volumes. Since the definition as well as all subsequent volume formulas all depend heavily on calculus, we shall operate entirely on the intuitive level. As with area, we will continue to make the same two basic assumptions on volume, namely,

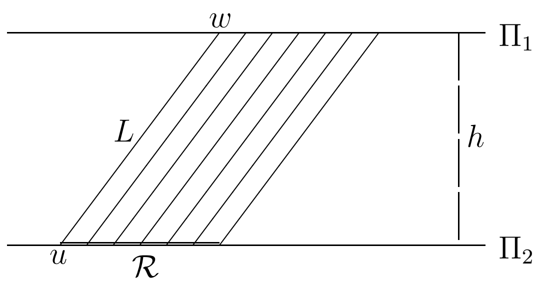
(i′) congruent solids have equal volume, and

(ii′) if a solid is divided into two sub-solids so that the latter intersect each other at most at their boundary surfaces, then the volume of the solid is the sum of the volumes of these two sub-solids.

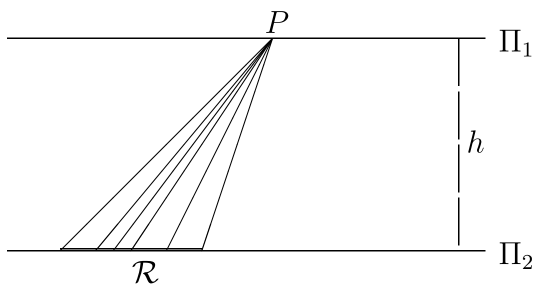
However, in contrast with the case of area where assumptions (i) and (ii) were used very effectively to derive basic area formulas for most of the common rectilinear planar regions, there is a mathematical reason, too advanced to be given here, why the same kind of reasoning with (i′) and (ii′) would not lead to similar volume formulas except in a few exceptions such as rectangular prisms (solids). Almost any volume formula must employ something equivalent to calculus. Therefore, the important thing at this point is to know how to correctly apply the formulas. For example, you should be able to decide whether a spherical container of radius 1.2 or a cubical container of side 2 contains more honey (*your* money may be at stake). It is also true that you owe it to yourself to learn, at least once in your life, *why* all these beautiful formulas are true.

The first formulas have to do with the volumes of "generalized pyramids" and "generalized cylinders". Let $\Pi_1$ and $\Pi_2$ be two planes in 3-space, parallel but of distance $h$

apart, and let $\mathcal{R}$ be a region in the plane $\Pi_2$. We first show how to generate generalized cylinders using $\mathcal{R}$ in $\Pi_2$. Let $w$ be a point in $\Pi_1$ and $u$ be a point in the given region $\mathcal{R}$ of $\Pi_2$. Let $L$ be the line segment joining $w$ to $u$. Now let $u$ trace out all the points of $\mathcal{R}$ and let $L$, while following $u$, remain parallel to itself. (The point $w$ therefore traces out a region in $\Pi_1$ congruent to $\mathcal{R}$ as $u$ sweeps out $\mathcal{R}$.) The solid $\mathcal{C}$ traced out by the line segment $L$ in this manner is called a *generalized cylinder* with base $\mathcal{R}$ and height $h$. (Thus $h$ is the distance between the top and bottom of $\mathcal{C}$.) Because it is difficult to draw three dimensional pictures, we draw a two-dimensional profile to give an idea:




In the same setting, we now describe how to generate a generalized pyramid with base $\mathcal{R}$. Fix a point $P$ in $\Pi_1$. Then the solid $\mathcal{V}$ which is the union of all the line segments obtained by joining $P$ to all the points in $\mathcal{R}$ is called a *generalized pyramid* with base $\mathcal{R}$ and height $h$. Again, we only draw a two-dimensional profile for illustration:



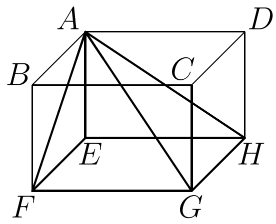

In case $\mathcal{R}$ is a disk and $L$ is perpendicular to $\Pi_2$ (and hence also to $\Pi_1$), the resulting generalized cylinder $\mathcal{C}$ is nothing but a *right circular cylinder* in the usual sense. If $\mathcal{R}$ is a disk, then the resulting generalized pyramid $\mathcal{V}$ is what is usually called a *cone*. In case $\mathcal{R}$ is a rectangle and $L$ is perpendicular to $\Pi_2$, $\mathcal{C}$ is a rectangular solid. In case $\mathcal{R}$ is a square and the point $P$ is right above the center of the square, the resulting $\mathcal{V}$ is then a *pyramid* in the usual sense.

The basic formulas in this context are:

$$\text{the volume of a generalized cylinder with base } \mathcal{R} \text{ and height } h = (\text{area of } \mathcal{R}) \cdot h. \tag{11}$$

$$\text{the volume of a generalized pyramid } \mathcal{V} \text{ with base } \mathcal{R} \text{ and height } h = \frac{1}{3}\,(\text{area of } \mathcal{R}) \cdot h. \tag{12}$$

Thus given a generalized pyramid and a generalized cylinder with the same base and height, the volume of the former is always $\dfrac{1}{3}$ of the latter. The proof of this fact requires calculus, but in an important special case, we can gain some insight into how the factor of $\frac{1}{3}$ arises. So let $\mathcal{C}_0$ be the unit cube $ABCDEFGH$ (i.e., all sides are of length 1) and let $\mathcal{V}_0$ be the generalized pyramid $AEFGH$ sharing the same base as the unit cube (see figure below). Thus the generalized cylinder ($\mathcal{C}_0$) and the generalized pyramid ($\mathcal{V}_0$) have the same square base and the same height (= 1).

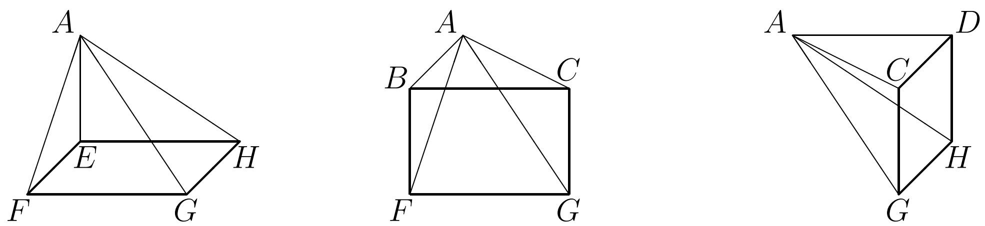



We now prove directly that

$$\text{volume of } \mathcal{V}_0 \;=\; \frac{1}{3}\,(\text{volume of } \mathcal{C}_0) \tag{13}$$

This is because the cube $\mathcal{C}_0$ is the union of the following three generalized pyramids:

$$AEFGH\ (= \mathcal{V}_0), \qquad ABFGC, \qquad \text{and} \qquad ADCGH.$$




These three generalized pyramids are congruent! The easiest way to see this is to realize that all three have congruent bases (unit squares), same height (= 1), and the vertex of each pyramid is directly above a vertex of the base ($AE$ is perpendicular to the base $EFGH$ on the left, $AB$ is perpendicular to the base $BFGC$ in the middle, and $AD$ is perpendicular to the base $DCGH$ on the right). Thus the unit cube is the union of three congruent pyramids which intersect only at boundary triangles ($AEFGH$ and $ABFGC$ intersect only at $\triangle AFG$, $ABFGC$ and $ADCGH$ intersect only at $\triangle AGC$, and $AEFGH$ and $ADCGH$ intersect only at $\triangle AGH$). By assumptions (i′), the volumes of all three pyramids are equal and therefore, by assumption (ii′), the volume of each such pyramid is a third of the volume of the unit cube.

The reason this special case is important is that, using basic properties of the behavior of volume under dilation in each coordinate direction and the so-called *Cavalieri's*

*Principle* (provable only with the help of calculus considerations), this trisection of the unit cube furnishes the basis of a proof of formula (12). See Chapter 9 in Serge Lang and Gene Murrow, **Geometry**, 2nd edition, Springer-Verlag, 1988.

We now go back to formulas (11) and (12). Using formula (7), we conclude from formula (11) that in case $\mathcal{R}$ is a disk of radius $r$, the following classical formulas are valid:

$$\text{the volume of a circular cylinder } \mathcal{R} \text{ of base radius } r \text{ and height } h$$
$$= \pi r^2 h. \tag{14}$$

$$\text{the volume of a cone of base radius } r \text{ and height } h = \frac{1}{3}\pi r^2 h. \tag{15}$$

Our final formula pertains to the familiar case of the ball:

$$\text{the volume of a ball of radius } r = \frac{4}{3}\pi r^3. \tag{16}$$

Formula (16) is one of the most celebrated formulas in the mathematics of antiquity. Like formula (10), it was also first discovered by Archimedes. We have more to say about him presently. The same formula was re-discovered by a different method, and most probably independently, several centuries later in China by Zu Chong-Zhi (430–501) and his son Zu Geng. Nowadays, you should be able to derive (16) in your first semester of calculus.

Now let us make some observations about the area formulas (7) and (10), and the volume formulas (14) and (16). First suppose we place a *hemisphere* (half a sphere) of radius $r$ on top of a disk of radius $r$. By formula (7), the area of the disk is $\pi r^2$, and by formula (10), the area of the hemisphere is $2\pi r^2$. Now compare the two areas, and you would suddenly be struck by the fact that the area of the hemisphere is exactly **twice** that of the disk. Would you have guessed that beforehand? Next, suppose we have a ball of radius $r$ *inscribed in a cylinder.* This means the top and bottom disks of the cylinder touch the sphere, as does the vertical wall of the cylinder. Notice that this means the radius of the cylinder is also $r$ and the height of the cylinder is $2r$. By formula (14), the volume of the cylinder is $2\pi r^3$. Comparing this with formula (16), we see that *the volume of a ball inscribed in a cylinder is* $\frac{2}{3}$ *the volume of the cylinder.* We go further: the surface area of the sphere which is the boundary of the ball under discussion is $4\pi r^2$, by formula (14). On the other hand, the (total) surface area of the cylinder is $\pi r^2 + \pi r^2 + (2\pi r)(2r)$ which is equal to $6\pi r^2$. Therefore *the surface area of a sphere inscribed in a cylinder is also* $\frac{2}{3}$ *the surface area of the cylinder.* These remarkable relationships between the surface areas and volumes of a ball inscribed in a cylinder were the discoveries of Archimedes. He was so proud of them that he requested that a picture of a ball inscribed in a cylinder be drawn on his tombstone.

A final thought: the volume of a rectangular solid with sides $a$, $b$, and $c$ is of course $abc$.

Here is a brief indication of why the concept of *area* is subtle. Let us agree that a *unit square* (i.e., each side has length 1) has area 1. It is then easy to convince oneself that (say) a rectangle of sides 3 and 7 must have area 21 (no matter how *area* is defined) because it can be neatly partitioned into exactly 21 unit squares. The same reasoning shows that if a rectangle has sides $m$ and $n$, where $m$ and $n$ are *integers*, then its area (again regardless of how it is defined) must be $mn$. What about a rectangle of sides 0.2 and 0.3? We can reason as follows: Let $Q$ be a square of side 0.1. Since a unit square can be partitioned into 100 copies of $Q$, the area of $Q$ is 1/100. Since also the rectangle of side 0.2 and 0.3 can be partitioned into 6 copies of $Q$, its area is therefore 6/100, which is of course just $0.2 \times 0.3$. The same reasoning then proves that a rectangle of sides $r$ and $s$, where $r$ and $s$ are *rational numbers*, has area equal to $rs$. So far so good. But suppose we have a rectangle $R$ whose sides are of length $\sqrt{2}$ and $\sqrt{3}$, can the above argument conclude that the area of $R$ must be $\sqrt{2} \times \sqrt{3}$? A little reflection would reveal that that argument now breaks down completely and that this new situation calls for an overhaul of what we mean by the *area* of $R$. It turns out that the proper definition of *area* must involve the concept of a *limit*, and that the new defintion (using limits) of area now leads to the desired conclusion that the area of a rectangle of sides $a$ and $b$ is $ab$, for any *real* numbers $a$ and $b$.

From this discussion, one can get a glimpse of the complications when one tries to define the *area* of a curved object like the sphere.