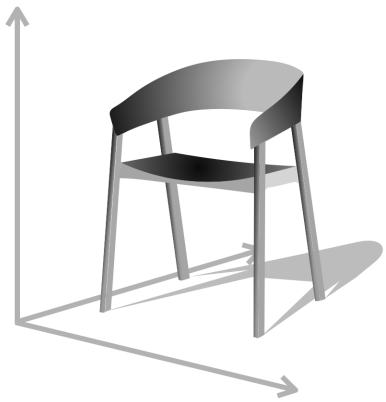
Dimensions in Mathematics

By Di Beos

Based on the Princeton Companion to Mathematics

What's the difference between a set which is two dimensional and one that is three dimensional? We say that a photo of a chair is a 2D representation of a 3 dimensional object because the actual chair has height, width, and depth, but it's photograph only shows us the height and width.



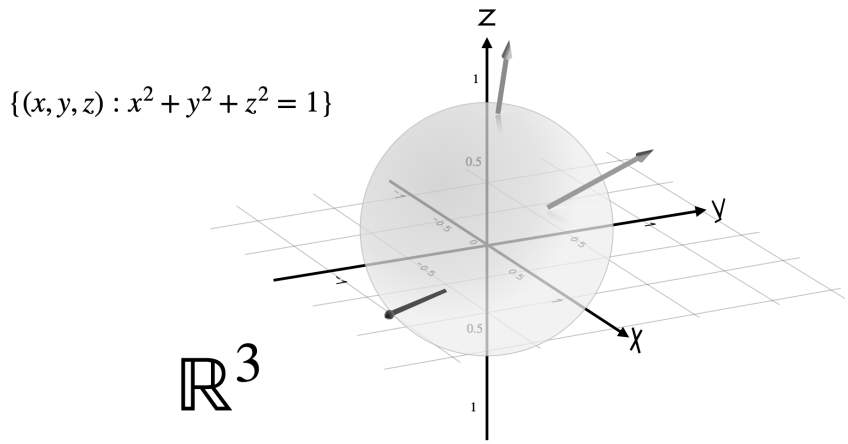
So the rough answer to our initial question is that a two dimensional set lives inside of a plane while a set that is three dimensional fills up a portion of space.

That seems to be the case for many sets, because triangles, squares, and circles can be easily drawn on a plane while shapes like tetrahedra, cubes, and spheres cannot be.

But what if we take something like the surface of a sphere? It looks to us two dimensional, much like a flat sheet of paper. But this contrasts with the actual solid sphere itself, which clearly takes three dimensions to describe it.

Would this therefore mean that the initial very rough definition was incorrect? Not exactly.

If we look at it from a linear algebra perspective, the set $\{(x, y, z) : x^2 + y^2 + z^2 = 1\}$ is the surface of a sphere of radius 1 in $\mathbb{R}^3$ centered at the origin. Although the sphere itself is an object embedded in $\mathbb{R}^3$, any point in $\mathbb{R}^3$ can be reached starting from the sphere and moving linearly in some direction. So the set is three dimensional because it cannot be contained in a plane.



But, this isn't very fair: the surface of a sphere, when considered as an entity in itself, does not occupy any volume. This example tells us something: the dimension of a set varies according to which definition we use, and they are often incompatible with one another.

One basic definition of dimension is "the number of coordinates one needs to specify a point". Physicists would call it the "degrees of freedom" of the system.

This can be used to justify the earlier understanding of the surface of a sphere being 2 dimensional. Any point can be specified by giving it a longitude and latitude. But this definition is a little tricky to make rigorous because you could technically define a point on the sphere using just *one* number.

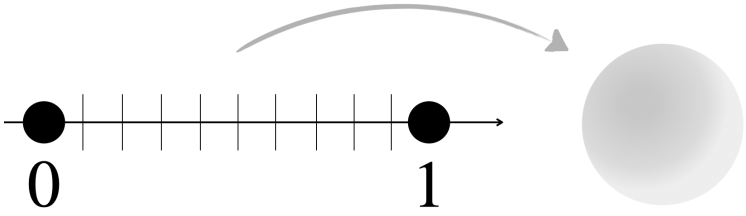
For example, take Pi and Euler's number.

- π = 3.141592653589793...
- e = 2.718281828459045...

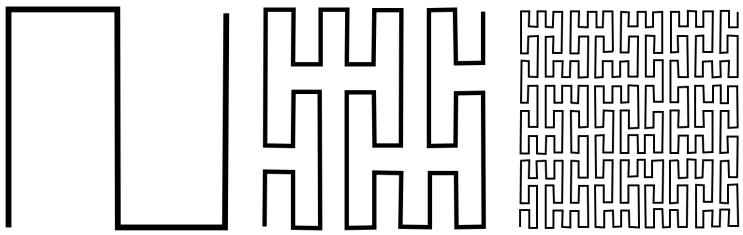
You can take these two and interleave the digits to form a single new number from which the two original numbers can be **recovered**. You start with the first digit of π, then the first digit of e, followed by the second digit of π, the second digit of e, and so on. This is of course really artificial, but it does reduce it to one number instead of 2.

But it's not only that that's the problem with our original definition. We can even find a continuous function f from the closed interval [0, 1] to the surface of a sphere. This interval represents all real numbers from 0 to 1, including both endpoints. It is a simple, one-dimensional line segment.

Basically what I'm saying is that we can have a single one-dimensional line segment [0,1] correspond to a two-dimensional surface (the sphere) in a way that covers every point on the sphere exactly once.



This involves space-filling curves, which are special types of continuous curves that, although one-dimensional, completely fill entire two-dimensional regions, like this Peano curve, first discovered by Giuseppe Peano in 1890, which is a continuous curve that passes through every point of a square.



Peano curve

In a sphere, you could theoretically design a continuous function that maps every number in the interval [0,1] to a distinct point on the sphere's surface. Such a function would effectively "unfold" the one-dimensional line into a two-dimensional surface in a continuous, smooth manner.

In order to solve this, we have to define what we mean by a "natural" coordinate system. This leads us to the definition of a manifold.

What we do is take our sphere, and select a local area which we will call *N*. We map this area, through the mapping ϕ, to correspond this area to a Euclidean space $\mathbb{R}^2$. This essentially allows us to treat this area of the sphere as flat, or 2 dimensional, and makes it much easier to deal with points and lines within that area.

If you want to know more about this in detail, check out these videos

- How to do Calculus on an Abstract Manifold
- How to Get to Manifolds Naturally

Thus at its core, the intuition that a d-dimensional set is one where d numbers are needed to specify a point can indeed be developed into a rigorous definition which will tell us that the surface of a **solid** sphere is two dimensional.

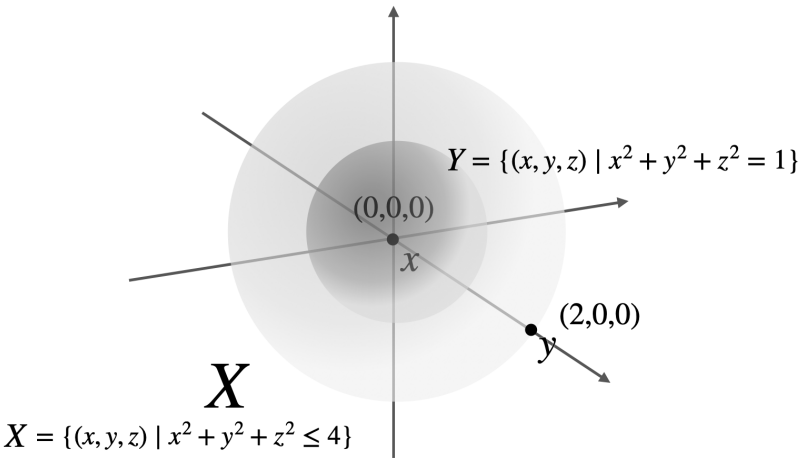
Let's say I want to cut a piece of paper into two pieces. The boundary of each cut piece will be a curve. We think of this curve as one dimensional. Why?

Because you only need one parameter to describe it (like "position along the line"). Let's use the same reasoning for this. Take the curve and cut it in half. The part where the two remaining pieces will meet each other will be a single point (or a pair of points if we're talking about a loop). This point is *zero dimensional*, because it has no length, width, or depth.

This creates the principle that **dividing a d-dimensional object creates a boundary of dimension (d−1).**

Let's say *d* is the number of dimensions of the plane, the sheet of paper, which is 2. The resulting curve has dimension d (which in this case is 2) minus 1. So the curve has dimension *d-1* or 1.

Say we have the set X, which is a solid sphere of radius 2, and it contains the points x and y. Let x be the center of the sphere (0, 0, 0), and y be a point on the boundary of the sphere, say (2, 0, 0). Now say there's a surface of a smaller sphere of radius 1 (centered at the origin), we'll call it Y. Y is something known as a barrier.

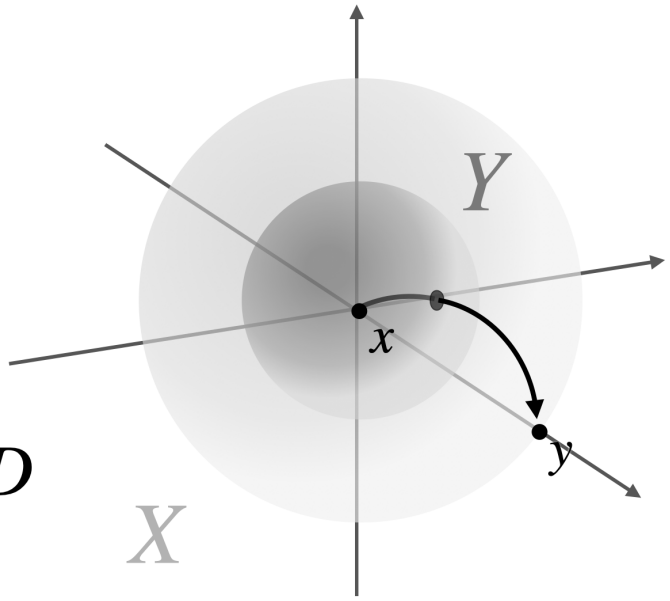


Any continuous path from the center (0, 0, 0) to the boundary point (2, 0, 0) must pass through Y. There is no way to "go around" the smaller sphere without intersecting it, because the entire path is confined to the solid sphere X.

Now, a **2D barrier** is sufficient to divide the solid sphere X, X is **at most 3D.** But, a 1D barrier (e.g., a curve) cannot divide X, it would leave the rest of the sphere connected. Thus, if we follow the *d-1* rule, X is not at most 2D, confirming that X is exactly 3D. Y which divides it has to be 2D, the line that connects x and y is 1D, and the point at which it is cut by the barrier is 0D.

Formally, X is at most *d-dimensional* if, between any two points in X, there exists a barrier of at most (d-1)-dimensional.

$$X \implies 3D$$
$$Y \implies 2D$$
$$\text{Line} \implies 1D$$
$$\text{Point} \implies 0D$$



A problem can arise with that definition of dimension when we construct a pathological set X that acts as a barrier between any two points in the plane, but contains no segment of any curve.

A pathological set is a set with strange, counterintuitive properties that challenge our usual geometric intuition. For example, imagine a set X that is so irregular that it acts as a barrier between any two points in the plane $\mathbb{R}^2$, meaning no continuous path can connect the points without passing through X.

By the definition, X would be considered a barrier, but it might not contain any segment of a curve (which we intuitively expect from a 1D object). Instead, X could be zero-dimensional (like a collection of points). However, if a zero-dimensional set like X can block paths across the plane, the definition implies that the plane is at most 1D, which is clearly wrong.

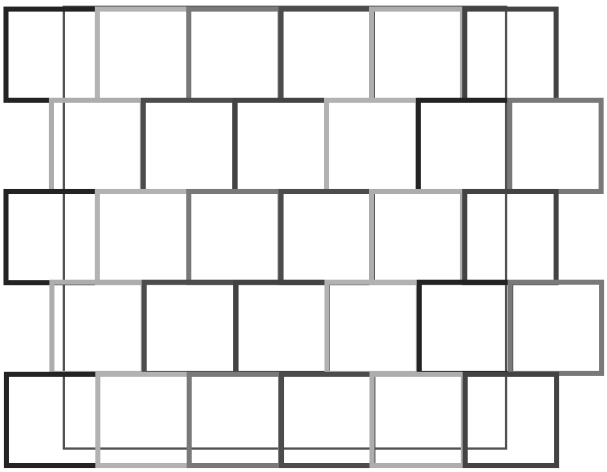
A **zero-dimensional set** (like a collection of points) could act as a barrier in a 2D space, falsely suggesting that the plane is only 1D.

But this is resolved with a small modification to the original definition that is called the *inductive dimension* of a set, introduced by Luitzen Brouwer.

This leads us to Lebesgue's **covering dimension.**

Covering a set means finding a collection of smaller sets that together "contain" the entire original set. For example, to cover an interval X=(0, 1), the open interval of real numbers (an interval that does not contain the endpoints 0 and 1), use a collection of smaller open intervals, like (0,0.4),(0.3,0.6),(0.5,0.9), and so on. These intervals "cover" X because every point in (0,1) lies inside at least one of these smaller intervals. The smaller intervals will **overlap** with each other to ensure no gaps, and you can do it in such a way that no point is contained in more than two of your intervals: just start each new interval close to the end of the previous one.

Now say we want to do that with a square, it's a set that does not contain the boundary of the square, and we cover it with other smaller squares. All of them will have to overlap, but this time, some of the points will have to overlap in at least three little squares right? If you stack them like this, you can do the covering in such a way that no four squares overlap.



The rule of thumb seems to be that to cover a d dimensional set, be it a line, square, cube, whatever, the overlaps are at least $d + 1$ and do not need to be greater than that.

Thus the precise definition is that a set X is at most d-dimensional if:

For any finite open cover $\{U_1, U_2, \ldots, U_n\}$ of X, you can refine the cover to a new finite open cover $\{V_1, V_2, \ldots, V_m\}$ such that:

1. $\{V_i\}$ also covers the whole set X

2. Each V_i is fully contained in at least one of the original U_i

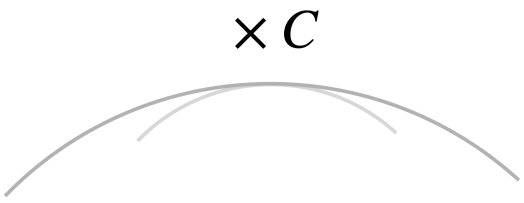
3. No point in X is contained in more than $d+1$ elements of the V_i

This definition is general enough to apply not only to subsets of $\mathbb{R}^n$ (like lines, squares, or cubes) but also to arbitrary topological spaces. The final idea about dimension will be how it affects how we measure *size*.

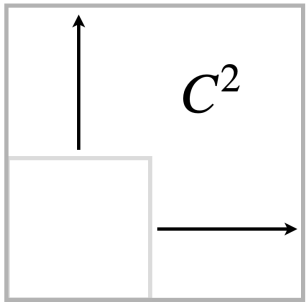
If we want to figure out how big X is, then if we give it length, it is one dimensional. If we give it area it is two dimensional. If we give it volume it is three dimensional. This assumes that you already know what the dimension is, you're not trying to figure it out. But, we'll actually see that there's a way of deciding which one it is without determining the dimension in advance.

The dimension can be defined to be the number that corresponds to the best measure. We will use the fact that length, area and volume scale in different ways when the shape is expanded.

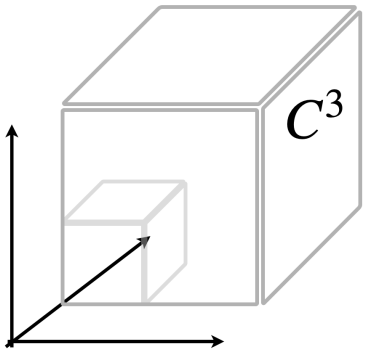
If we take a curve and expand it by a factor of 2, in all directions, it's length doubles. Generally speaking if we expand it by C it grows by C.



If we take a shape which is 2 dimensional and expand it by C it multiplies by C^2. This is generally speaking, because each portion of the shape expands in "two directions", so the area has to be multiplied by C twice.



When it is a 3D object, it multiplies by C^3 because it scales in three directions, and so on.



It may still appear as though we have to decide in advance whether length, area, or volume will be talked about, before understanding how it will scale. But that's actually not the case.

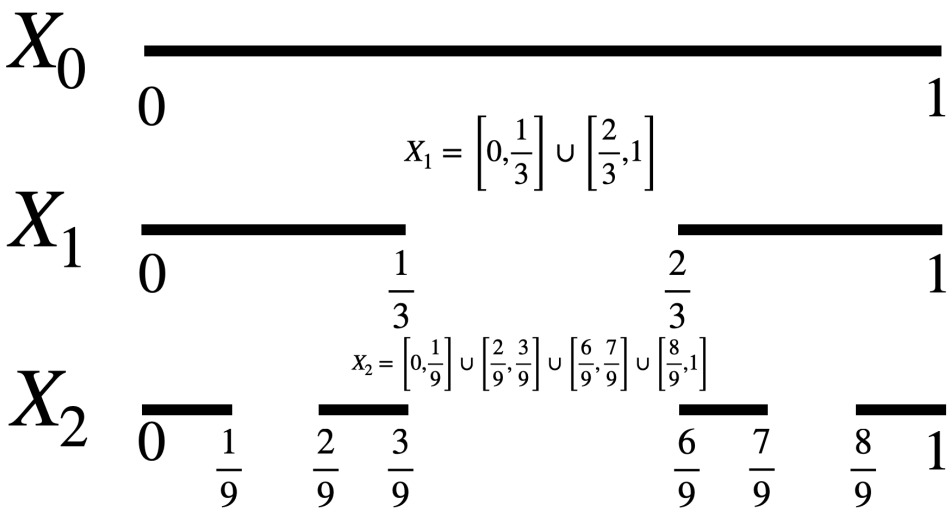
If we take a square and scale it by a factor of 2, the new square can be divided into **four smaller squares**, each congruent to the original square. The side length doubles (in two directions: horizontally and vertically), so the area becomes $2^2 = 4$ times the original area.

Scaling behavior reveals the dimensionality of an object without needing to decide in advance whether we are measuring its length, area, or volume. This has a very interesting consequence: there are sets to which it is natural to assign a dimension that is not an integer!

The simplest example is a Cantor set.

Start with a closed interval (so including the endpoints) [0,1], call it X_0. Then we form a set X_1 by removing the middle third of X_0, so all the points between $\frac{1}{3}$ and $\frac{2}{3}$, leaving $\frac{1}{3}$ and $\frac{2}{3}$ themselves. X_1 is therefore the union of the closed intervals [0, $\frac{1}{3}$] and [$\frac{2}{3}$, 1]. Next, we remove

the middle thirds of these two closed intervals to produce a set X_2. So, X_2 is this. And so on, repeated indefinitely.



Generally speaking, X_n is a union of closed intervals. X_{n+1} is what you get when you remove the middle thirds of each interval. Basically, X_{n+1} consists of twice as many intervals as X_n, but they are a third of the size.

Once the sequence X_0, X_1, X_2,... is produced, you define the Cantor set to be the intersection of all the X_i: that is, all the real numbers that remain, no matter how far you go with the process of removing middle thirds of intervals.

To figure out the **dimension** of the Cantor set, we analyze its scaling behavior using the formula for fractal dimension d: $N = C^d$.

N is the number of "pieces" (intervals). C is the scaling factor (how much smaller each piece becomes). If you go on to explore what the actual dimension is, it brings us to The Hausdorff dimension of the Cantor set, which is equal to ln(2)/ln(3) ≈ 0.631.

The point is that, by analyzing how the number of pieces and their size change during scaling, we can compute dimension, even for sets that are highly fragmented like the Cantor set.

Please, if you find this document useful, let us know. Or if you found typos and things to improve, let us know as well. Your feedback is very important to us. We're working hard to deliver the best material possible. Contact us at: dibeos.contact@gmail.com