CARDINALITIES[1]

The sets A and B are called *equinumerous* if there is a one-to-one function $f : A \to B$ onto B. Writing $\mathbb{N} = \{1, 2, 3, \dots\}$, we say that A is *countably infinity* if $\mathbb{N}$ and A are equinumerous. A is said to be *countable* if A is finite or countably infinite.

Lemma 1. *If $X \subset \mathbb{N}$ is infinite then X is countably infinite.*

Proof. Define the function $f : X \to \mathbb{N}$ as follows. For each $n \in X$ put

$$f(n) = \min\{i \in \mathbb{N} : i \neq f(k) \text{ for } k < n \text{ with } k \in X\}.$$

Observe that this is a recursive definition; that is, the definition of $f(n)$ relies on the definition of $f(k)$ for $k < n$ with $k \in X$.[2]

We claim that f is one-to-one. In fact, if $n, k \in X$ and $k < n$, then the definition explicitly asserts that $f(n) \neq f(k)$. Further, we claim that f is onto $\mathbb{N}$. In fact, assume, on the contrary, that, for some $m \in \mathbb{N}$, there is no $k \in X$ such that $f(k) = m$. Then, for every $n \in X$ we have

$$m \in \{i \in \mathbb{N} : i \neq f(k) \text{ for } k < n \text{ with } k \in X\}.$$

As $f(n)$ is the least element of the set on the right-hand side, this implies that $f(n) \leq m$. That is, $f(n) \leq m$ for every $n \in X$. Since f is one-to-one and X is infinite, this is not possible.[3] □

Corollary. *Assume A is countably infinite and $B \subseteq A$ is infinite. Then B is countably infinite.*

Proof. Let $f : \mathbb{N} \to A$ be a one-to-one function onto A. Put

$$X = \{n \in \mathbb{N} : f(n) \in B\}.$$

Then X is infinite, so it is countably infinite by the above lemma. Let $g : \mathbb{N} \to X$ be a one-to-one function onto X. Then the function $f \circ g : \mathbb{N} \to A$ is[4] one-to-one and onto B, showing that B is countably infinite. □

Lemma 2. *The set $\mathbb{Z}$ of all integers is countably infinite.*

Proof. For a real number x, write $\lfloor x \rfloor$ for the largest integer $\leq x$. The function $f : \mathbb{N} \to \mathbb{Z}$ defined by

$$f(n) = (-1)^n \left\lfloor \frac{n}{2} \right\rfloor \qquad (n \in \mathbb{N})$$

is one-to-one and onto $\mathbb{Z}$. Indeed, we have $f(1) = 0$, for $k \in \mathbb{N}$ we have $f(2k) = k$ and $f(2k + 1) = -k$. □

---

[1] Notes for Course Mathematics 9.5 at Brooklyn College of CUNY. Attila Máté, April 21, 2009.

[2] One might reflect that this definition gives $f(n) = 1$ for the least element of n of X, in which case the restriction on i after the colon is vacuous, since there is no $k < 1$ with $k \in X$; that is, the clause after the colon is true for every $i \in \mathbb{N}$ in this case.

[3] If we take m to be the least integer for which no $k \in X$ exists such that $f(k) = m$, then we can in fact conclude by this argument that the range of the function f is the set $\{1, 2, \dots, m - 1\}$; i.e., that X has exactly $m - 1$ elements.

[4] We could have written $f \circ g : \mathbb{N} \to B$ instead of $f \circ g : \mathbb{N} \to A$, since it is easy to verify that all values of $f \circ g$ are in B. The emphasis here, however, is on the word "verify"; however easy the verification of this fact is, to see that the values of $f \circ g$ are in A can be seen much more directly, since all values of f are in A.

Note. We have

$$\left\lfloor \frac{n}{2} \right\rfloor = \frac{n}{2} + \frac{(-1)^n - 1}{4}.$$

Indeed, for even n, the right-hand side gives $n/2$, and for odd n it gives $(n-1)/2$. Hence we can also define the above function f as

$$f(n) = (-1)^n \frac{2n + (-1)^n - 1}{4} \qquad (n \in \mathbb{N}).$$

Lemma 3. *The Cartesian product $\mathbb{N} \times \mathbb{N}$ is countably infinite.*

Proof. A one-to-one function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ can be defined as follows. Given $(m, n) \in \mathbb{N} \times \mathbb{N}$, put[5]

$$f(m, n) = \frac{(m + n - 1)(m + n)}{2} + n - 1$$

for $m, n \in \mathbb{N}$.

We claim that f is one-to-one. To show this, let (m, n) and (k, l) be two pairs of positive integers such that $(m, n) \neq (k, l)$.[6] Without loss of generality, we may assume that $m + n \leq k + l$. If $m + n = k + l$ then we must have $n \neq l$, so

$$\begin{aligned} f(m, n) &= \frac{(m + n - 1)(m + n)}{2} + n - 1 = \frac{(k + l - 1)(k + l)}{2} + n - 1 \\ &\neq \frac{(k + l - 1)(k + l)}{2} + l - 1 = f(k, l). \end{aligned}$$

If $m + n < k + l$ then we have $m + n \leq k + l - 1$ and $m + n + 1 \leq k + l$, and so

$$\begin{aligned} f(m, n) &= \frac{(m + n - 1)(m + n)}{2} + n - 1 < \frac{(m + n - 1)(m + n)}{2} + m + n - 1 \\ &= \frac{(m + n - 1)(m + n) + 2(m + n)}{2} - 1 = \frac{(m + n + 1)(m + n)}{2} - 1 \\ &= \frac{(m + n)(m + n + 1)}{2} - 1 \leq \frac{(k + l - 1)(k + l)}{2} - 1 \\ &< \frac{(k + l - 1)(k + l)}{2} + l - 1 = f(k, l), \end{aligned}$$

so again $f(m, n) \neq f(k, l)$. This shows that f is one-to-one.

Write

$$X = \{n \in \mathbb{N} : n = f(k, l) \text{ for some } k, l \in \mathbb{N}\}.$$

Then $X \subseteq \mathbb{N}$ is infinite, and so X is countably infinite. Since $f : \mathbb{N} \times \mathbb{N} \to X$ is a one-to-one mapping that is onto X, this shows that $\mathbb{N} \times \mathbb{N}$ is also countably infinite, which is what we wanted to prove. □

Note. The function f defined in the last proof is not onto $\mathbb{N}$. The function $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ defined by

$$g(m, n) = \frac{(m + n - 2)(m + n - 1)}{2} + n$$

for $m, n \in \mathbb{N}$ is onto $\mathbb{N}$. The calculations showing that g is one-to-one are slightly more complicated than the ones showing that f is one-to-one, and showing that g is onto $\mathbb{N}$ requires extra effort. Since it was not important in the above proof that f be onto $\mathbb{N}$, it was simpler to work with the function f instead of g.

[5] Strictly speaking, an element of $\mathbb{N} \times \mathbb{N}$ has the form (m, n) for positive integers m and n. The value of f at such an element should be denoted as $f\big((m, n)\big)$. However, it visually more pleasing to use the notation $f(m, n)$ at the price of some inaccuracy.

[6] That is, $m \neq n$ or $k \neq l$.

Corollary. *If A and B are countably infinite sets then $A \times B$ is also countably infinite.*

Proof. Let $f : \mathbb{N} \to A$ onto A, $g : \mathbb{N} \to B$ onto B, and $h : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ onto $\mathbb{N} \times \mathbb{N}$ be one-to-one functions. Such functions exist since A and B are countably infinite by assumption, and $\mathbb{N} \times \mathbb{N}$ is countably infinite by the last lemma.

We define the function $\phi : \mathbb{N} \to A \times B$ as follows. Given $n \in \mathbb{N}$, let $k, l \in \mathbb{N}$ be such that $h(n) = (k, l)$, and let $\phi(n) = (f(k), g(l))$. It is easy to show that ϕ is one-to-one and onto $A \times B$. $\square$

In what follows, $\mathbb{Q}$ will denote the set of rational numbers, and $\mathbb{Q}^+$ will denote the set of positive rationals. We have

Lemma 4. *$\mathbb{Q}^+$ is countably infinite.*

Proof. Write

$$S = \{(m, n) : m, n \in \mathbb{N} \text{ and the greatest common divisor of } m \text{ and } n \text{ is } 1\}.$$

The set S is an infinite subset of the countably infinite set $\mathbb{N} \times \mathbb{N}$, so it is countably infinite by the Corollary to Lemma 1. The function f defined as

$$f(m, n) = \frac{m}{n} \qquad \text{for} \quad (m, n) \in S$$

is a one-to-one function from S onto $\mathbb{Q}^+$, showing that $\mathbb{Q}^+$ is also countably infinite. $\square$

For a set A, the power set $\mathcal{P}(A)$ of A is defined as the set of all subsets of A. The following theorem and its proof is valid for any set A, be A finite or infinite; the proof is valid even when A is the empty set.[7] But only the case when A is infinite is of real interest, since for finite A a much more precise statement can be made.

Theorem. *Let A be an arbitrary set. Then A and $\mathcal{P}(A)$ are not equinumerous.*

Proof. We will show that there is no one-to-one function from A onto $\mathcal{P}(A)$; in fact, no function from A onto $\mathcal{P}(A)$ exist, whether or not we require it to be one-to-one. To see this, let $f : A \to \mathcal{P}(A)$ be an arbitrary function, and consider the subset C of A defined as

$$C = \{x \in A : x \notin f(x)\}.$$

Then there is no $y \in A$ for which $f(y) = C$. Indeed, for an arbitrary $y \in A$, if $y \in f(y)$ then $y \notin C$ by the definition of C, and if $y \notin f(y)$ then $y \in C$. This shows that $f(y)$ and C do not have the same elements (y is an element of exactly one of these two sets), so $f(y) \neq C$, as claimed. $\square$

A related argument can be given that $\mathbb{N}$ and the set of real numbers, $\mathbb{R}$, are not equinumerous. On the other hand, we have the following

Theorem. *The set $\mathbb{R}$ and the interval $(-1, 1)$ are equinumerous.*

Proof. The function $f : (-1, 1) \to \mathbb{R}$ such that

$$f(x) = \frac{x}{x^2 - 1} \qquad \text{for} \quad x \in (-1, 1)$$

is one-to-one and onto $\mathbb{R}$. Indeed, let $y \in \mathbb{R}$ be arbitrary. We need to show that there is exactly one $x \in (-1, 1)$ such that $f(x) = y$. If $y = 0$ then we have $f(x) = y$ only for $x = 0$. Assume now that $y \neq 0$. Then the equation $f(x) = y$ can be equivalently written as $y(x^2 - 1) = x$.

Observe that this latter equation makes sense for $x = \pm 1$ while the equation $f(x) = y$ does not. The important point, however, is that $x = \pm 1$ does not satisfy the latter equation, since for this choice of x the left-hand side is 0 while the right-hand side is ± 1. That is, the exceptional case of $x = \pm 1$ does not affect the equivalence of the two equations.

[7] As one might expect, the case when A is the empty set involves a number of vacuously true statements.

Keeping in mind that we assumed that $y \neq 0$, the latter equation can also be written as

$$x^2 - \frac{1}{y}x - 1 = 0.$$

This is a quadratic equation for x. We can solve this equation for x as

$$x = \frac{\frac{1}{y} \pm \sqrt{\frac{1}{y^2} + 4}}{2}.$$

Given that the discriminant (the expression under the square root) of this equation is positive, this equation has two distinct real solutions; call them x_1 and x_2. The product of these two solutions is the constant term of the equation; that is, $x_1 x_2 = -1$. Therefore $|x_1||x_2| = 1$. Given that $|x_1|, |x_2| \neq 1$, as we remarked above, one of x_1 and x_2 must be inside the interval $(-1, 1)$ and the other one must be outside. That is, there is exactly one $x \in (-1, 1)$ for which $f(x) = y$, as we wanted to show. $\square$

THE CANTOR-SCHRÖDER-BERNSTEIN THEOREM

Consider the sets $A = \mathbb{R}$ and $B = [-1, 1]$, the closed interval from -1 to 1. There there are functions $f : A \to B$ into B and $g : B \to A$ into A that are one-to-one. Namely, A is equinumerous to $(-1, 1)$ according to the last theorem; so we can take f to be the one-to-one function from A onto $(-1, 1)$. For the function g we can simply take the identity function on B. Since $B \subset A$, g is into A. The next theorem asserts that under these conditions A and B are equinumerous.

Cantor-Schröder-Bernstein Theorem. *Let A and B be sets and assume there are one-to-one functions $f : A \to B$ and $g : B \to A$. Then A and B are equinumerous.*

In the theorem, it is of course not required that f be onto B or g be onto A; in fact, there would be nothing to prove if this were the case. The result is intuitively obvious if A or B are finite sets, but for infinite sets the result is not at all obvious, and a proof is needed. The proof, however, works regardless whether A or B are finite or infinite. We will give two proofs. The first proof gives a much better insight why the result is true, especially if one follows along the first proof by drawing a picture. The second proof can be presented more concisely, but it gives little insight why the result is true. The second proof is the one that is usually given in textbooks. Before giving the formal version of the first proof, we include an intuitive description.

Imagine the sets A and B as two vertical lines, A on the left, B on the right. From each point, left or right, draw one, and if possible, two edges, one forward edge to the image of the point under the function f or g (whichever is applicable), and one backward edge, to the the inverse image. The forward image always exists, the backward image may not exist, since the inverses need not be defined everywhere. These edges can be continued to complete paths containing the given point. Two different paths may not have points in common, since the functions f and g are one-to-one. A path may have a starting point, where the inverse image does not exist, or may go back indefinitely (when it may loop back on itself, forming a cycle, or may not; a cycle will always contain an even number of edges). Select alternate edges of these paths, making sure that the starting point of the path is incident to an edge. (When the path has no starting point, it is immaterial how the edges are selected, but below we select the edge going backward from left to right, since this slightly simplifies the formal description). The selected edges form a one-to-one mapping from A to B (when redirected from left to right if necessary). The mapping is defined on all of A and is onto B for the same reason: each point is incident to a path.

We will now give a formal description of the proof.

First Proof. For a function ϕ one usually denotes by $\phi(x)$ the value of ϕ at x, but one sometimes uses the notation ϕx instead. It will be advantageous for us to use this latter notation in what follows in order to

avoid having to write too many parentheses.[8] Denote by f^{-1} and g^{-1} the inverses of the functions f and g. For every $a \in A$, form the sequence

$$a, \quad g^{-1}a, \quad f^{-1}g^{-1}a, \quad g^{-1}f^{-1}g^{-1}a, \quad f^{-1}g^{-1}f^{-1}g^{-1}a, \ \ldots .$$

Note that g^{-1} maps a subset of A into B, and f^{-1} maps a subset of B into A. Hence $g^{-1}a$ may or may not be defined. Even if $g^{-1}a \in B$ is defined $f^{-1}g^{-1}a$ may not be defined, and if $f^{-1}g^{-1}a \in A$ is defined, the element $g^{-1}f^{-1}g^{-1}a$ may then not be defined. That is, the above sequence may terminate at some point. Call the sequence associated with an $a \in A$ in this way $\sigma(a)$.

The elements of this sequence need not be distinct; for example, we can have $f^{-1}g^{-1}a = a$, in which case the sequence never terminates, but it has only two distinct elements, a and $g^{-1}a$. When talking about the *number of elements* of this sequence, we will mean its length, and not the number of its distinct elements. For example, in case $f^{-1}g^{-1}a = a$ we will say that the above sequence has infinitely many elements, even though the number of its distinct elements is only two.

Now, define a function $h : A \to B$ as follows. For an arbitrary $a \in A$, consider the above sequence. If (i) the sequence $\sigma(a)$ has infinitely many elements, or a finite even number of elements, then put $ha = g^{-1}a$, and if (ii) the $\sigma(a)$ has an odd number of elements then put $ha = fa$. First, observe that this defines ha for every $a \in A$. Indeed, in case (i), the sequence $\sigma(a)$ has at least two elements, so $g^{-1}a$ is defined; so ha is defined in case (i); since fa is defined for every $a \in A$, ha is also defined in case (ii).

We show that h is one-to-one. To this end, let $a_1, a_2 \in A$ such that $a_1 \neq a_2$. If both ha_1 and ha_2 are defined according to clause (i), then $ha_1 = g^{-1}a_1 \neq g^{-1}a_2 = ha_2$. Similarly, if both ha_1 and ha_2 are defined according to clause (ii), then $ha_1 = fa_1 \neq fa_2 = ha_2$. So assume one of ha_1 and ha_2 is defined according to clause (i), and the other according to clause (ii). Without loss of generality, we may assume that ha_1 is defined according to clause (i) and ha_2 is defined according to clause (ii). Assume, that $ha_1 = ha_2$, i.e., $g^{-1}a_1 = fa_2$. Then $a_2 = f^{-1}g^{-1}a_1$. Hence the sequence $\sigma(a_1)$ can be written as

$$a_1, \quad g^{-1}a_1, \quad \sigma_1(a_2), \quad \sigma_2(a_2), \quad \sigma_3(a_2), \ \ldots ,$$

where $\sigma_1(a_2)$, $\sigma_2(a_2)$, $\sigma_3(a_2)$, $\ldots$ denote the first, second, third, $\ldots$ elements of the sequence $\sigma(a_2)$. Now, $\sigma(a_2)$ has an odd number of elements, since clause (ii) was used to define ha_2, and $\sigma(a_1)$ has either an infinite number of elements or a finite even number of elements, since clause (i) was used to define ha_1. This is a contradiction, since we just saw that $\sigma(a_1)$ has exactly two more elements than $\sigma(a_2)$. This contradiction shows that h is one-to-one.

To show that h is onto B, let $b \in B$ be arbitrary, and define the sequence

$$b, \quad f^{-1}b, \quad g^{-1}f^{-1}b, \quad f^{-1}g^{-1}f^{-1}b, \quad g^{-1}f^{-1}g^{-1}f^{-1}b, \ \ldots .$$

Call this sequence $\rho(b)$. Let $a_1 = gb$. Then $b = g^{-1}a_1$, and so the sequence $\sigma(a_1)$ can be written as

$$a_1, \quad \rho_1(b), \quad \rho_2(b), \quad \rho_3(b), \ \ldots ,$$

where $\rho_1(b)$, $\rho_2(b)$, $\rho_3(b)$, $\ldots$ denote the first, second, third, $\ldots$ elements of the sequence $\rho(b)$. That is, $\sigma(a_1)$ has one more elements than $\rho(b)$. Hence, if $\rho(b)$ has an infinite number of elements or a finite odd number of elements then $\sigma(a_1)$ has either an infinite number of elements or a finite even number of elements, and so $ha_1 = g^{-1}a_1 = b$.

Assume now that $\rho(b)$ has a finite even number of elements. Then it has at least two elements, so $a_2 = f^{-1}b$ is defined. Then the elements of the sequence $\rho(b)$ can be written as

$$b, \quad \sigma_1(a_2), \quad \sigma_2(a_2), \quad \sigma_3(a_2), \ \ldots ,$$

showing that $\sigma(a_2)$ has one fewer element than $\rho(b)$. That is, $\sigma(a_2)$ has an odd number of elements, and so $ha_2 = fa_2 = b$. This shows that h is onto B. $\square$

[8] The notation ϕx could be confusing where *juxtaposition* (i.e., placing next to each other) of letters can indicate multiplication; this is why one usually uses the notation $\phi(x)$ instead. In the present case, multiplication is not used, so there is no such danger.

The second proof defines the same one-to-one function $h : A \to B$, but it describes this function in a different way. After carefully reading both proofs, one should realize that the two proofs are essentially the same, presented differently.

Second Proof. Write $A_0 = A$, $B_0 = B$, and if A_n and B_n for $n \geq 0$ have been defined, put

$$A_{n+1} = \{g(b) : b \in B_n\} \quad \text{and} \quad B_{n+1} = \{f(a) : a \in A_n\}.$$

Note that since f is one-to-one, this means that, for every $n \geq 0$, $a \in A_n$ if and only if $f(a) \in B_{n+1}$.[9] Similarly, $b \in B_n$ is and only if $g(b) \in A_{n+1}$.

We clearly have $A_1 \subseteq A_0$ and $B_1 \subseteq B_0$. By induction, it is then easy to prove that $A_{n+1} \subseteq A_n$ and $B_{n+1} \subseteq B_n$. Indeed, if we assume that for some $n > 1$ we have both $A_n \subseteq A_{n-1}$ and $B_n \subseteq B_{n-1}$, then $A_{n+1} \subseteq A_n$ follows from the latter of these relations, and $B_{n+1} \subseteq B_n$ follows from the former. Let

$$C = \bigcap_{n=0}^{\infty} A_n.$$

Define the function h on A as follows. If (i) $a \in A_n \setminus A_{n+1}$ for some odd n or $a \in C$ then put $h(a) = g^{-1}(a)$, where g^{-1} is the inverse of g, and if (ii) $a \in A_n \setminus A_{n+1}$ for some even n, then put $h(a) = f(a)$. We then have to show that (1) the definition of h is meaningful, (2) h is defined for every $a \in A$, (3) h is one-to-one, and (4) h is onto B.

As for (1) and (2), let $a \in A$. Then either $a \in C$ or there is an k for which $a \notin A_k$. Assume that $a \notin C$, and let $k \geq 0$ be the least integer for which $a \notin A_k$. Then $a \in A_m$ for every nonnegative integer $m < k$ and $a \notin A_m$ for every $m \geq k$. Hence for the integer n such that $a \in A_n \setminus A_{n+1}$ we must have $n = k - 1$. That is, this integer n is uniquely determined. This makes the definitions given in (i) and (ii) meaningful, except that we need to show that $g^{-1}(a)$ is defined in case (i). However, $g^{-1}(a)$ is defined unless $a \in A_0 \setminus A_1$, and $a \in A_0 \setminus A_1$ is not true in case (i).

As for (3), assume $h(a_1) = h(a_2)$ for some $a_1, a_2 \in A$ such that $a_1 \neq a_2$. Since f and g^{-1} are one-to-one, this is not possible if both $h(a_1)$ and $h(a_2)$ are both defined by clause (i) or if they are both defined by clause (ii). Assume, therefore, that they are defined by different clauses. Without loss of generality, we may assume that $h(a_1)$ is defined by clause (i) and $h(a_2)$ is defined by clause (ii). We then have $g^{-1}(a_1) = h(a_1) = h(a_2) = f(a_2)$. Since we used clause (ii) to define $h(a_2)$, we have $a_2 \in A_n \setminus A_{n+1}$ for some even n. Writing $b = f(a_2)$, we have $b \in B_{n+1} \setminus B_{n+2}$. Since we also have $b = g^{-1}(a_1)$, i.e., $a_1 = g(b)$, it follows that $a_1 \in A_{n+2} \setminus A_{n+3}$. As $n + 2$ is even, this contradicts the assumption that $h(a_1)$ was defined according to clause (i).

As to (4), let $b \in B$. If $b \in B_n$ for all $n \geq 0$ then, writing $a = g(b)$, we have $a \in A_n$ for all $n \geq 0$, and so $a \in C$. Therefore, $h(a) = g^{-1}(a) = b$ according to clause (i). Assume that this is not the case, and let $n \geq 0$ be the unique integer such that $b \in B_n \setminus B_{n+1}$. If n is even then, with $a = g(b)$, we have $a \in A_{n+1} \setminus A_{n+2}$. Since $n + 1$ is odd, we have $h(a) = g^{-1}(a) = b$ according to clause (i). If n is odd, then $n \geq 1$; thus $b \in B_n \subseteq B_1$, which implies that $b = f(a)$ for some $a \in A$. Then $a \in A_{n-1} \setminus A_n$, and $h(a) = f(a) = b$ according to clause (ii). This completes the proof. $\square$

[9] The important point here is that if $a \notin A_n$ then $f(a) \notin B_{n+1}$. Indeed, if we had $f(a) \in B_{n+1}$, there would have to be an $a' \in A_n$ for which $f(a') = f(a)$. But this is not possible, since f is one-to-one, and so we would have to have $a' = a$, but $a \notin A_n$.