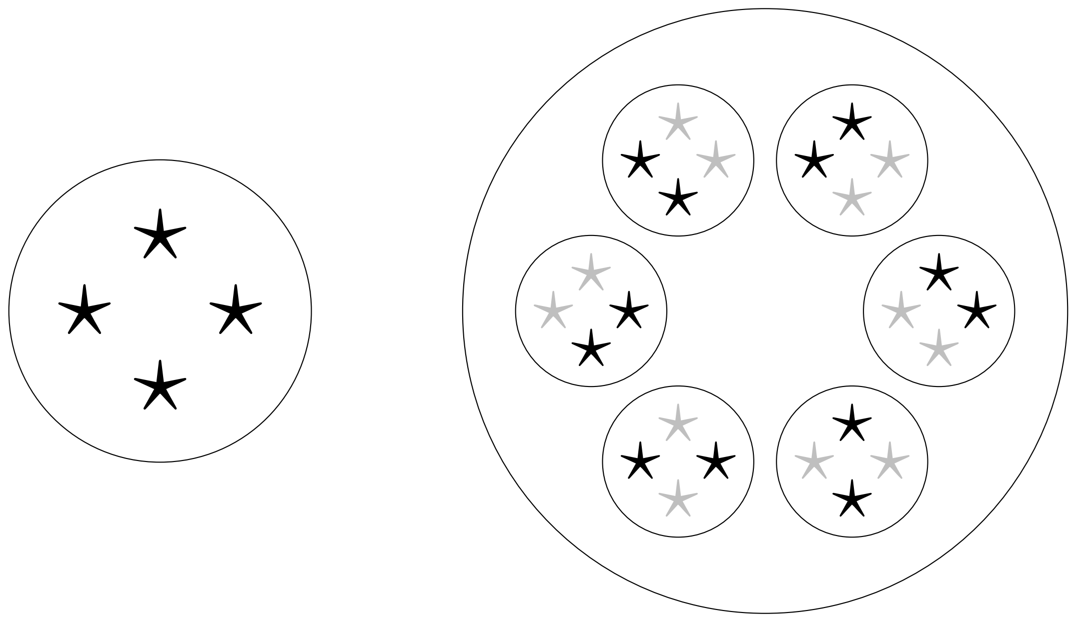
Combinatorics is the study of counting, so numbers generally represent the "size" of a set of objects. For example, the number 4 could represent the quantity of stars in the left-hand circle. As a slightly more complicated example, the number 6 could represent all the different ways to pick two out of these four stars.

Definition. A **set** is any collection of "things" (we will only be concerned with finite sets, though we will not rigorously define the word "finite"). These "things" in A are called **elements** of A, and we write $a \in A$ to mean "a is an element of A." We write $\#(A)$ to denote the number of elements of A.

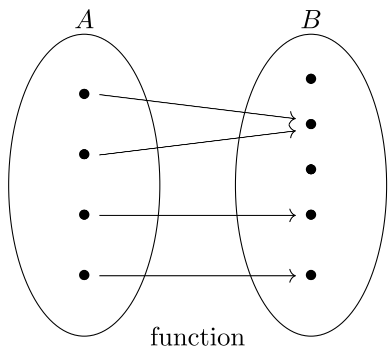


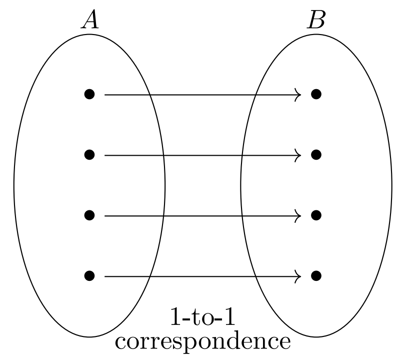
When doing math, we often get equations like $2 \times 6 = 3 \times 4$. Since we're used to the rules of arithmetic, we can do such calculations purely symbolically, but *what do they mean*? Each side represents a quantity, the number of things in some set. The equality means that these two sets have the same number of things. For example, we can interpret $2 \times 6 = 3 \times 4$ as meaning that a 2-by-6 grid and a 3-by-4 grid have the same number of 1-by-1 squares. Before proceeding any further, we will make this reasoning a little more formal.

Intuitive Definition. A **1-to-1 correspondence** is a rule that pairs elements of sets A and B, such that each element of A corresponds to exactly one element of B and vice versa. When counting the size of a set, the assertion $\#(A) = n$ can be taken to mean "there is a 1-to-1 correspondence between A and $\{1, 2, \dots, n\}$."

We can **compose** 1-to-1 correspondences $A \longleftrightarrow B$ and $B \longleftrightarrow C$ to get a 1-to-1 correspondence $A \longleftrightarrow C$. Therefore $\#(A) = n = \#(B)$ if and only if there is a 1-to-1 correspondence $A \longleftrightarrow \{1, 2, , \dots, n\} \longleftrightarrow B$.

Formal Definition. Given sets A and B, a **function** $f : A \to B$ is a rule that takes in any element $a \in A$ and assigns it an element $f(a) \in B$. One example is the **identity** $\mathrm{Id}_A : A \to A$, which is defined by the rule $f(a) = a$ for all $a \in A$. Two functions $f : A \to B$ and $g : B \to C$ **compose** to form a function $g \circ f : A \to C$, which is defined by $(g \circ f)(a) = g\big(f(a)\big)$ for all $a \in A$ (notice the order in which we write the two functions). Two functions $f : A \to B$ and $g : B \to A$ are said to be **inverses** if they satisfy $g \circ f = \mathrm{Id}_A$ and $f \circ g = \mathrm{Id}_B$. A function $f : A \to B$ is called a **1-to-1 correspondence** if it admits an inverse function $g : B \to A$.





Consider natural numbers $n \geq k \geq 0$. If we have a set of n objects, the number of ways to pick out k objects is denoted $\binom{n}{k}$ (pronounced n-choose-k). You may have seen this before, with the concrete formula

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

(Recall the definition of the factorial $n! = 1 \times 2 \times \cdots \times n$.) However, our goal is not calculate the concrete number of ways, but rather to set up 1-to-1 correspondences that show equality without needing computation.

Examples. We will demonstrate some examples of bijective proofs using this "choose" operation.

(a) Show that $\binom{n}{k} = \binom{n}{n-k}$ for any $0 \leq k \leq n$.

Let A be the set of ways to pick a committee[1] of k people out of a larger group of n people.

Let B be the set of ways to pick a committee of $n-k$ people out of a larger group of n people.

For every choice $a \in A$ of a committee of k people, we can get a committee $f(a) \in B$ of $n-k$ people by choosing everyone *not* in A. This is a 1-to-1 correspondence between A and B, so we have

$$\binom{n}{k} = \#(A) = \#(B) = \binom{n}{n-k}.$$

Usually, we won't spell it out in this much detail, but rather say something like "to choose a committee of k people from a total of n people, it is equivalent to choose the $n-k$ people not in the committee."

(b) Show that $\binom{n}{k}\binom{k}{j} = \binom{n}{j}\binom{n-j}{k-j}$ for any $0 \leq j \leq k \leq n$.

The quantity $\binom{n}{k}\binom{k}{j}$ is the number of ways to pick a committee of k people out of a total of n people, then pick a sub-committee of j people out of those k. It is equivalent to pick a j-person sub-committee out of the larger group of n people, then pick the remaining $k-j$ committee members from the $n-j$ people remaining. This latter process can be conducted in $\binom{n}{j}\binom{n-j}{k-j}$ different ways.

(c) $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$ for any $1 \leq k \leq n-1$.

There are $\binom{n}{k}$ ways to select a k-person committee from a larger group of n people, of whom I am one. Consider whether or not I am chosen. If I am not, then there are $\binom{n-1}{k}$ ways to choose the committee members from among the other $n-1$ people. But if I am chosen, then there are $\binom{n-1}{k-1}$ ways to choose the remaining members of the committee. Therefore, in total, there are $\binom{n-1}{k} + \binom{n-1}{k-1}$ possibilities.

[1] Thinking in terms of groups of people, rather than arbitrary sets, can make these problems more concrete. In this setting, "committee" is the typical term for a distinguished subset of the total group of people.

Exercises. Here are some more identities that admit bijective proofs.

(a) Show that $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$ for any $n \geq 0$. Next, prove the following generalization:

$$\sum_{k=j}^{n} \binom{n}{k}\binom{k}{j} = 2^{n-j}\binom{n}{j} \qquad \text{for any } 0 \leq j \leq n.$$

(b) Show that $\binom{k}{k} + \binom{k+1}{k} + \cdots + \binom{n}{k} = \binom{n+1}{k+1}$ for any $0 \leq k \leq n$ (this is called the hockey-stick formula). Next, prove the following generalization, which is called the Chu-Vandermonde identity:

$$\sum_{j=0}^{k} \binom{m}{k}\binom{n}{k-j} = \binom{m+n}{k}.$$

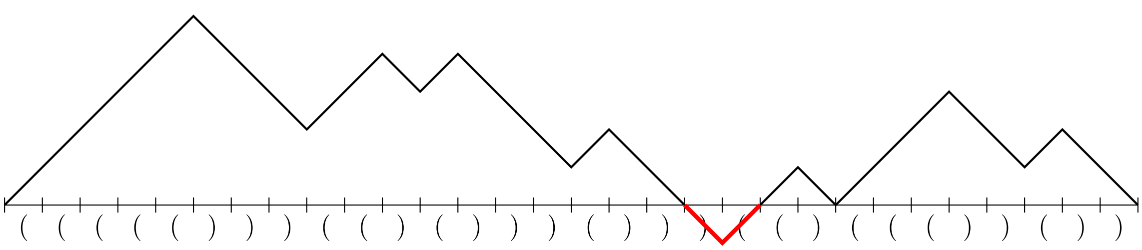
(Here, we must define $\binom{n}{k} = 0$ whenever $k > n$. Why does this definition make sense?)

Hopefully, these examples have acquainted us with the idea of bijective proofs. We now move on to our main class of examples, the **Catalan numbers**. These are a infinite sequence of numbers $C_0, C_1, C_2, C_3, \ldots$, which has many different, equivalent definitions. We have to start somewhere, so we define them as follows:

Definition. A **Dyck word** of length $2n$ is a valid arrangement of n pairs of parentheses. As a first example, when $n = 2$, there are two valid arrangements: ()() and (()); as well as four invalid arrangements:)((),)()(,))((and ())(. More formally, an arrangement is valid if, at any point along the sequence, there are no more closing parentheses than opening parentheses. We define C_n to be the number of Dyck words of length $2n$.

Examples. We saw that $C_2 = 2$. It is easy to check that $C_1 = 1$. We also have $C_0 = 1$, because we count the "empty word" as one valid arrangement of 0 pairs of parentheses. See if you can calculate C_3 and C_4.

As the number of parentheses grows, checking the validity of a Dyck word becomes a strain on the eyes. Is ((((()))(()()))()))(()((())()) valid? It's hard to tell! To make this easier, we can instead draw **Dyck paths**. Drawing $2n$ steps along a flat line segment (the ground), we can write a "step up" instead of an opening parenthesis and a "step down" instead of a closing parenthesis. This will look something like the following:

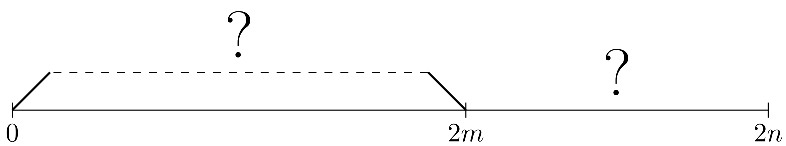


Let us now revisit our conditions in terms of this "mountain range." Our elevation at a given point is

$$(\#\text{ of opening parentheses}) - (\#\text{ of closing parentheses}).$$

Thus, the fact that parentheses come in pairs means that we start and end at ground-level. The fact that there are no more closing than opening parentheses means that the above difference should always be ≥ 0, so we can never dip below ground-level. In particular, we now *easily* see where ((((()))(()()))()))(()((())()) violates the validity condition (note the red valley). In essence, we have established a 1-to-1 correspondence between Dyck words and Dyck paths. Now, let's try to actually count them!

Consider a Dyck path of length $2n$. It may dip back down to ground-level somwhere between the beginning and ending of the path, but this must happen after an even number of steps (after an odd number of steps, our elevation will be odd and thus non-zero). So let us count the Dyck paths that first touch down after $2m$ steps, where $0 < m \leq n$ (if $m = n$, then we don't touch back down until the end). This will look as follows:



Between $2m$ and $2n$, we can have any Dyck path of length of $2(n-m)$. But in the first part, we cannot dip below the dashed line, or we would touch down somewhere before $2m$. Hence, by shifting our view so that this dashed line corresponds to elevation 0, we see that this portion can be any Dyck path of length $2(m-1)$ (since it goes from 1 to $2m-1$). In other words, the number of Dyck paths that first touch back down after $2m$ steps is $C_{m-1}C_{n-m}$. Summing over all of the possible values $m = 1, 2, \ldots, n$, we get a recursion formula:

$$C_n = \sum_{m=1}^{n} C_{m-1}C_{n-m}.$$

Starting with $C_0 = 1$, this allows us to compute any Catalan number from the previous ones. But it would be nicer to have a *closed form*, where we can calculate C_n without first having to calculate C_m for all $m < n$. We will prove the following formula, which allows for much simpler calculations:

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

To do so, we must allow for paths that may dip below ground-level. Thus, we say "path" to mean a sequence of up/down steps which starts and ends at ground-level, but which may dip below somewhere in-between. Thus, the only requirement is that there are an equal number of up-steps and down-steps. Therefore, a path is determined by choosing which n of the $2n$ steps are up-steps. Hence, we can count a total of $\binom{2n}{n}$ paths.

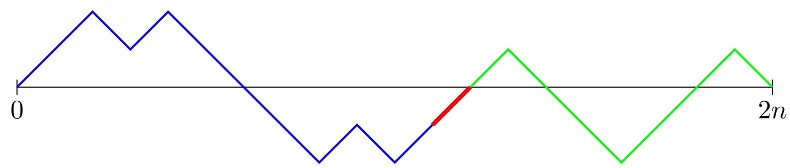
Exercise. Translate this back into the language of parentheses. If we randomly pick out some arrangement of n opening and n closing parentheses, what is the probability that this arrangement is valid?

We will say that the deficit of a path is k if there are exactly $2k$ steps where the path is below ground-level (by an argument similar to the one above, this number of steps must be even). Then there are $n+1$ possible values for the deficit, namely $k = 0, 1, \ldots, n$. We will show that there are actually an equal number of paths for each of these given deficits. In particular, the number of Dyck paths (i.e. paths of deficit 0) is $\frac{1}{n+1}$ times the total number of paths, which we saw to be $\binom{2n}{n}$. This will immediately yield the formula $C_n = \frac{1}{n+1}\binom{2n}{n}$.

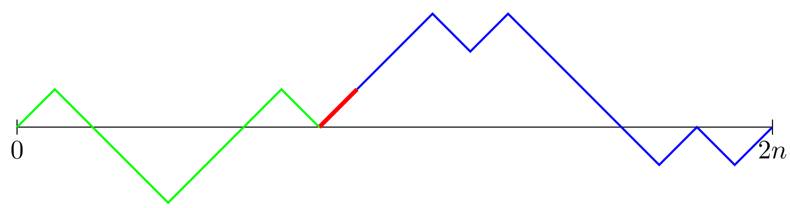
Proof that there are the same number of paths for each deficit. It suffices to fix $1 \leq k \leq n$ and find a 1-to-1 correspondence between paths of deficit k and paths of deficit $k-1$ (for we can then iterate this process). Suppose that we have a path of deficit k. We will transform it to have deficit $k-1$. Since $k > 0$, it must dip below ground-level. Consider the first two steps going between elevations 0 and -1 (either up or down):

- The first such step is a down-step and is the first time we dip below ground-level;
- The second such step is an up-step and is the first time we return to ground-level from below.

Mark this second step red, everything before it blue and everything after it green. (See the figure below.)



Move the green portion to the beginning and the blue portion to the end. Move the red step to fit between the newly positioned green and blue pieces. (See the figure below.) Since the green portion starts and ends at elevation 0, its portion of the deficit in the green portion remains unchanged. But the blue portion ends at elevation -1, while a path is required to end at elevation 0, so we need to shift the blue portion up by 1. There is only one blue step between elevations 0 and -1, since we colored the second such step red. Thus, shifting the blue portion up by 1 only moves 1 blue step from below-ground to above-ground. The red step also moves from below-ground to above-ground, since it now goes between the end of green (at elevation 0) and the beginning of blue (at elevation 1). Hence, two steps are moved from below-ground to above-ground, one red and one blue. This shows that we indeed reduced the deficit by 1.



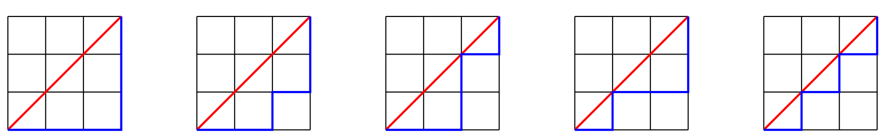
Lastly, to show that this gives a 1-to-1 correspondence, we must find an inverse operation going from paths of deficit $k-1$ to paths of deficit k. Given a path of deficit $k-1 < k \leq n$, there must be some portion of the path that is above-ground. Define the red step to be the last up-step between elevations 0 and 1, and then run an analogous procedure to the above. The details are similar, so we do not repeat them here. However, it is important to confirm that the two functions are inverses, which we have not done! The reader may attempt to fill in these details. (For a slicker definition of the second function, we can do the following: rotate the path by 180°, perform our original operation that reduces deficit by 1, then rotate back.) $\square$

Notice that, even when we are concretely counting sets, we are still able to implement bijective proofs: though Dyck paths are hard to count directly, we considered them as part of a larger set that we could count, then used a bijective proof to see what fraction of this set they made up.

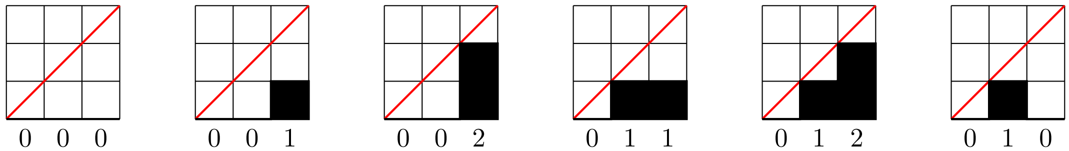
Another interpretation of Catalan numbers occurs when we rotate Dyck paths. Consider an $n \times n$ grid. Consider the set of all paths from the bottom-left to the top-right corner, staying below the main diagonal, where we can only travel up or to the right along grid-lines. (All such paths for $n = 3$ are drawn below.) Notice that we must take n steps to the right and n steps up, and staying below the diagonal means that

$$(\# \text{ of steps to the right}) - (\# \text{ of steps up})$$

is always ≥ 0. Thus, the number of such paths is C_n. (In fact, these can just be seen as rotated Dyck paths!)



Let L_n denote the set of sequences $(a_1, \ldots, a_n)$ with $0 \le a_k < k$ for all $k = 1, \ldots, n$. There are k choices for the value of a_k, so $\#(L_n) = 1 \times 2 \times \cdots \times n = n!$. We can illustrate such a sequence with a bar graph:



The condition "$0 \le a_k < k$ for all $k = 1, \ldots, n$" means precisely that the bar graph lies below the diagonal. We will call such a sequence **non-decreasing** if $a_1 \le a_2 \le \cdots \le a_k$. Given any such sequence, we can start in the bottom-left corner and travel up and to right along the bar graph to the top-left corner. This yields one of the grid-paths lying below the diagonal, which we discussed above. Conversely, given any such path, we get a bar graph of a non-decreasing sequence by filling in everything below the path. (Compare the two figures above.) This gives a 1-to-1 correspondence between non-decreasing sequences in L_n and Dyck paths.

Shifting gears a little bit, we now return to the concept of a function.

Definition. For two sets A and B, we say that A is a **subset** of B (written $A \subset B$) if every element of A is also in B (i.e. $a \in A \implies a \in B$). If $f : B \to C$ is a function and $A \subset B$ is a subset, we define a subset $f(A) \subset C$ to consist of the elements $f(a) \in C$ for all $a \in A$ (this is called the **image** of A).

Consider a function $f : A \to B$. We will say that …

- … f is onto if every element of B is hit by *at least* one element of A (i.e. $f(A) = B$),
- … f is 1-to-1 if every element of B is hit by *at most* one element of A (i.e. $f(a_1) = f(a_2) \implies a_1 = a_2$),
- … f is a 1-to-1 correspondence if every element of B is hit by *exactly* one element of A.

This definition of a 1-to-1 correspondence is just a slight rephrasing of the intuitive definition given above. Clearly, the function f is a 1-to-1 correspondence if and only if f is onto and 1-to-1. (The interested reader may prove that this definition of a 1-to-1 correspondence is equivalent to the definition in terms of inverses.)

Proposition. If $f : B \to C$ is 1-to-1 and $A \subset B$, then f restricts to a 1-to-1 correspondence $A \longleftrightarrow f(A)$.

Proof. We consider the "restricted" function $f|_A : A \to C$, where we only care about inputs coming from A. Since $f : B \to C$ is 1-to-1, the restriction $f|_A : A \to C$ is also 1-to-1. This function $A \to C$ may not be onto, but if we shrink the "target set" down to $f(A)$, we still get a well-defined function $f|_A : A \to f(A)$, since

$$a \in A \implies f|_A(a) = f(a) \in f(A).$$

The function $f|_A : A \to f(A)$ remains 1-to-1, but now it is also onto, simply by definition of $f(A)$. $\square$