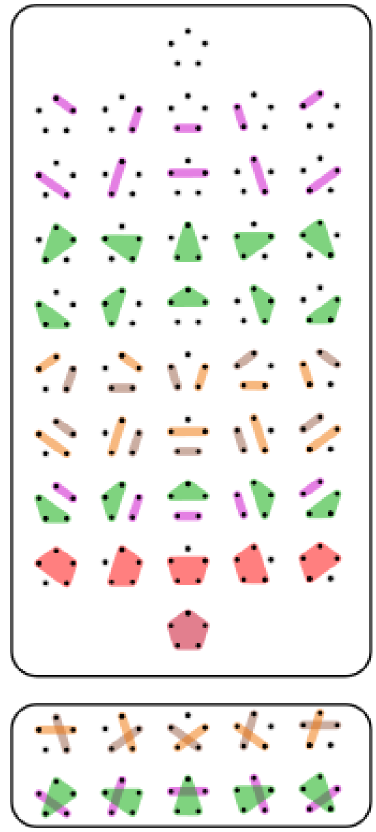
WIKIPEDIA
The Free Encyclopedia

Catalan number

In combinatorial mathematics, the **Catalan numbers** are a sequence of natural numbers that occur in various counting problems, often involving recursively defined objects. They are named after the French-Belgian mathematician Eugène Charles Catalan.



The $C_5 = 42$ noncrossing partitions of a 5-element set (below, the other 10 of the 52 partitions)

The nth Catalan number can be expressed directly in terms of the central binomial coefficients by

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!\,n!} = \prod_{k=2}^{n} \frac{n+k}{k} \qquad \text{for } n \geq 0.$$

The first Catalan numbers for n = 0, 1, 2, 3, ... are

1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, ...
(sequence A000108 in the OEIS).

Properties

An alternative expression for C_n is

$$C_n = \binom{2n}{n} - \binom{2n}{n+1} \text{ for } n \geq 0,$$

which is equivalent to the expression given above because $\binom{2n}{n+1} = \frac{n}{n+1}\binom{2n}{n}$. This expression shows that C_n is an integer, which is not immediately obvious from the first formula given. This expression forms the basis for a proof of the correctness of the formula.

Another alternative expression is

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

which can be directly interpreted in terms of the cycle lemma; see below.

The Catalan numbers satisfy the recurrence relations

$$C_0 = 1 \quad \text{and} \quad C_n = \sum_{i=1}^{n} C_{i-1} C_{n-i} \quad \text{for } n > 0$$

and

$$C_0 = 1 \quad \text{and} \quad C_n = \frac{2(2n-1)}{n+1} C_{n-1} \quad \text{for } n > 0.$$

Asymptotically, the Catalan numbers grow as

$$C_n \sim \frac{4^n}{n^{3/2}\sqrt{\pi}},$$

in the sense that the quotient of the nth Catalan number and the expression on the right tends towards 1 as n approaches infinity. A more accurate asymptotic analysis shows that the Catalan numbers are approximated by the fourth order approximation

$$C_n \sim \frac{4^{n-5}(8n(16n(8n-9)+145)-1155)}{\sqrt{\pi}n^{9/2}}$$

.

This can be proved by using the asymptotic growth of the central binomial coefficients, by Stirling's approximation for $n!$, or via generating functions.

The only Catalan numbers C_n that are odd are those for which $n = 2^k - 1$; all others are even. The only prime Catalan numbers are $C_2 = 2$ and $C_3 = 5$.[1]

The Catalan numbers have the integral representations

$$C_n = \frac{1}{2\pi}\int_0^4 x^n \sqrt{\frac{4-x}{x}}\,dx = \frac{2}{\pi}4^n \int_{-1}^1 t^{2n}\sqrt{1-t^2}\,dt.$$

which immediately yields $\sum_{n=0}^{\infty} \frac{C_n}{4^n} = 2$.

This has a simple probabilistic interpretation. Consider a random walk on the integer line, starting at 0. Let -1 be a "trap" state, such that if the walker arrives at -1, it will remain there. The walker can arrive at the trap state at times 1, 3, 5, 7..., and the number of ways the walker can arrive at the trap state at time $2k+1$ is C_k. Since the 1D random walk is recurrent, the probability that the walker eventually arrives at -1 is $\sum_{n=0}^{\infty} \frac{C_n}{2^{2n+1}} = 1$.

Applications in combinatorics

There are many counting problems in combinatorics whose solution is given by the Catalan numbers. The book *Enumerative Combinatorics: Volume 2* by combinatorialist Richard P. Stanley contains a set of exercises which describe 66 different interpretations of the Catalan numbers. Following are some examples, with illustrations of the cases $C_3 = 5$ and $C_4 = 14$.

- C_n is the number of Dyck words[2] of length $2n$. A Dyck word is a string consisting of n X's and n Y's such that no initial segment of the string has more Y's than X's. For example, the following are the Dyck words up to length 6:

$$\begin{gathered}\text{XY}\\ \text{XXYY} \quad \text{XYXY}\\ \text{XXXYYY} \quad \text{XYXXYY} \quad \text{XYXYXY} \quad \text{XXYYXY} \quad \text{XXYXYY}\end{gathered}$$

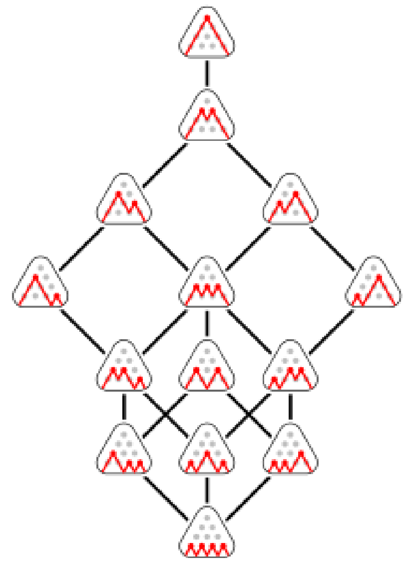
- Re-interpreting the symbol X as an open parenthesis and Y as a close parenthesis, C_n counts the number of expressions containing n pairs of parentheses which are correctly matched:

$$((())) \quad (()()) \quad (())() \quad ()(()) \quad ()()()$$

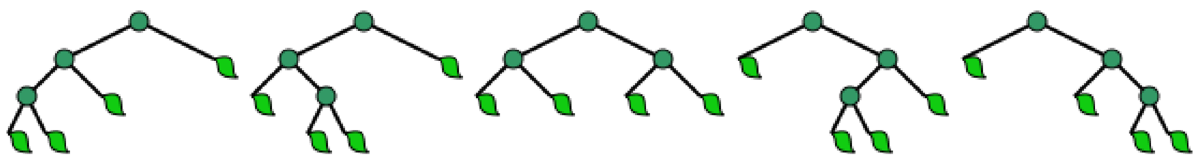
- C_n is the number of different ways $n + 1$ factors can be completely parenthesized (or the number of ways of associating n applications of a binary operator, as in the matrix chain multiplication problem). For $n = 3$, for example, we have the following five different parenthesizations of four factors:

$$((ab)c)d \quad (a(bc))d \quad (ab)(cd) \quad a((bc)d) \quad a(b(cd))$$

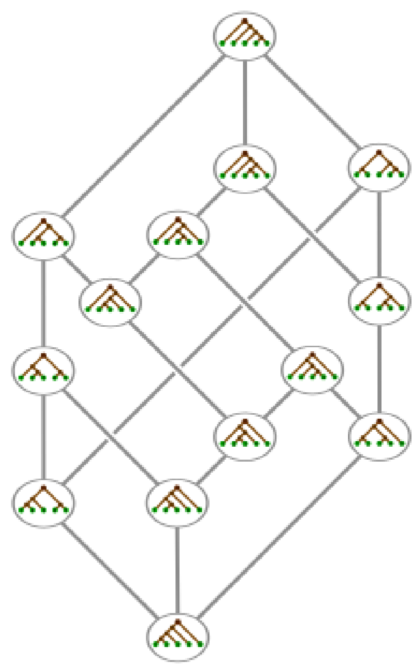
- Successive applications of a binary operator can be represented in terms of a full binary tree, by labeling each leaf *a,b,c,d*. It follows that C_n is the number of full binary trees with $n + 1$ leaves, or, equivalently, with a total of n internal nodes:



Lattice of the 14 Dyck words of length 8 – *(* and *)* interpreted as *up* and *down*

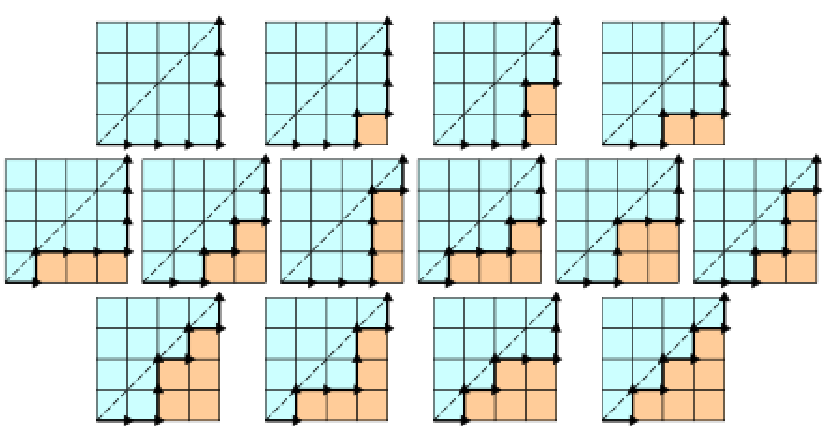


- C_n is the number of non-isomorphic ordered (or plane) trees with $n + 1$ vertices.[3] See encoding general trees as binary trees. For example, C_n is the number of possible parse trees for a sentence (assuming binary branching), in natural language processing.
- C_n is the number of monotonic lattice paths along the edges of a grid with $n \times n$ square cells, which do not pass above the diagonal. A monotonic path is one which starts in the lower left corner, finishes in the upper right corner, and consists entirely of edges pointing rightwards or upwards. Counting such paths is equivalent to counting Dyck words: X stands for "move right" and Y stands for "move up".



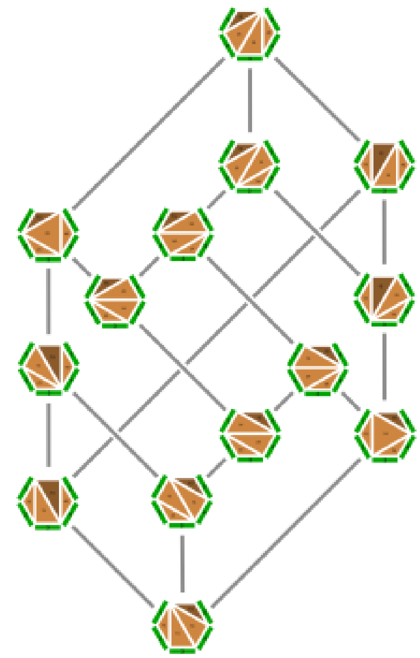
The associahedron of order 4 with the C_4=14 full binary trees with 5 leaves

The following diagrams show the case $n = 4$:



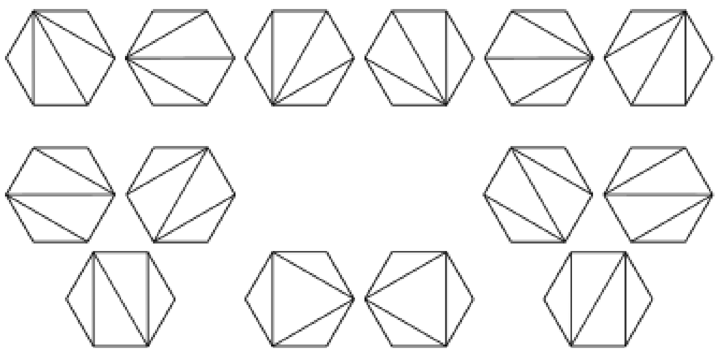
This can be represented by listing the Catalan elements by column height:[4]

[0,0,0,0] [0,0,0,1] [0,0,0,2] [0,0,1,1]
[0,1,1,1] [0,0,1,2] [0,0,0,3] [0,1,1,2] [0,0,2,2] [0,0,1,3]
[0,0,2,3] [0,1,1,3] [0,1,2,2] [0,1,2,3]

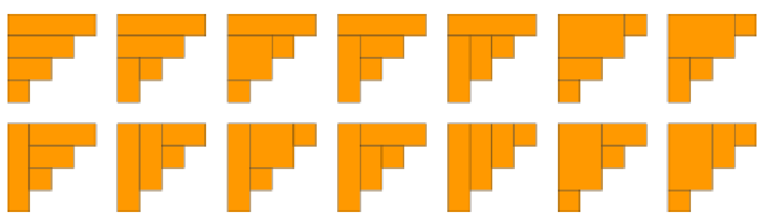


The dark triangle is the root node, the light triangles correspond to internal nodes of the binary trees, and the green bars are the leaves.

- A convex polygon with $n + 2$ sides can be cut into triangles by connecting vertices with non-crossing line segments (a form of polygon triangulation). The number of triangles formed is n and the number of different ways that this can be achieved is C_n. The following hexagons illustrate the case $n = 4$:



- C_n is the number of stack-sortable permutations of {1, ..., n}. A permutation w is called stack-sortable if $S(w) = (1, ..., n)$, where $S(w)$ is defined recursively as follows: write $w = unv$ where n is the largest element in w and u and v are shorter sequences, and set $S(w) = S(u)S(v)n$, with S being the identity for one-element sequences.
- C_n is the number of permutations of {1, ..., n} that avoid the permutation pattern 123 (or, alternatively, any of the other patterns of length 3); that is, the number of permutations with no three-term increasing subsequence. For $n = 3$, these permutations are 132, 213, 231, 312 and 321. For $n = 4$, they are 1432, 2143, 2413, 2431, 3142, 3214, 3241, 3412, 3421, 4132, 4213, 4231, 4312 and 4321.
- C_n is the number of noncrossing partitions of the set {1, ..., n}. *A fortiori*, C_n never exceeds the nth Bell number. C_n is also the number of noncrossing partitions of the set {1, ..., $2n$} in which every block is of size 2.
- C_n is the number of ways to tile a stairstep shape of height n with n rectangles. Cutting across the anti-diagonal and looking at only the edges gives full binary trees. The following figure illustrates the case $n = 4$:



- C_n is the number of ways to form a "mountain range" with n upstrokes and n downstrokes that all stay above a horizontal line. The mountain range interpretation is that the mountains will never go below the horizon.

Mountain Ranges

$n = 0:$	*	1 way
$n = 1:$	/\	1 way
$n = 2:$	/\ /\/\, / \	2 ways
$n = 3:$	/\ /\ /\ /\/\ / \ /\/\/\, /\/ \, / \/\, / \, / \	5 ways

- C_n is the number of standard Young tableaux whose diagram is a 2-by-n rectangle. In other words, it is the number of ways the numbers 1, 2, ..., $2n$ can be arranged in a 2-by-n rectangle so that each row and each column is increasing. As such, the formula can be derived as a special case of the hook-length formula.

```
123   124   125   134   135
456   356   346   256   246
```

- C_n is the number of length n sequences that start with 1, and can increase by either 0 or 1, or decrease by any number (to at least 1). For $n = 4$ these are $1234, 1233, 1232, 1231, 1223, 1222, 1221, 1212, 1211, 1123, 1122, 1121, 1112, 1111$. From a Dyck path, start a counter at *0*. An X increases the counter by *1* and a Y decreases it by *1*. Record the values at only the X's. Compared to the similar representation of the Bell numbers, only 1213 is missing.

Proof of the formula

There are several ways of explaining why the formula

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

solves the combinatorial problems listed above. The first proof below uses a generating function. The other proofs are examples of bijective proofs; they involve literally counting a collection of some kind of object to arrive at the correct formula.

First proof

We first observe that all of the combinatorial problems listed above satisfy Segner's[5] recurrence relation

$$C_0 = 1 \quad \text{and} \quad C_{n+1} = \sum_{i=0}^{n} C_i\, C_{n-i} \quad \text{for } n \geq 0.$$

For example, every Dyck word w of length ≥ 2 can be written in a unique way in the form

$w = \mathrm{X}w_1\mathrm{Y}w_2$

with (possibly empty) Dyck words w_1 and w_2.

The generating function for the Catalan numbers is defined by

$$c(x) = \sum_{n=0}^{\infty} C_n x^n.$$

The recurrence relation given above can then be summarized in generating function form by the relation

$$c(x) = 1 + xc(x)^2;$$

in other words, this equation follows from the recurrence relation by expanding both sides into power series. On the one hand, the recurrence relation uniquely determines the Catalan numbers; on the other hand, interpreting $xc^2 - c + 1 = 0$ as a quadratic equation of c and using the quadratic formula, the generating function relation can be algebraically solved to yield two solution possibilities