Indexing Combinations

Wagner Truppel

December 5, 2019

The problem

Suppose we're given the set $S_n = \{1, 2, \ldots, n\}$ of $n > 0$ integers and construct all combinations of these integers taken k at a time, where $0 < k \leq n$. Of course, there are

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

such combinations for any given pair of values (n, k). Constructing such combinations in an efficient manner is possible but is not a trivial exercise. It would be nice if we could build an indexing system in the form of an invertible function that maps any particular combination to a unique index, and vice-versa.

For that to work, however, we'd first have to establish an order between combinations. If the elements in a combination are arranged in ascending order, then two such combinations can be ordered by the order of the first non-matching pair of values, one value from each combination. Thus, for example, if $n = 8$ and $k = 4$, then $\{1, 3, 4, 8\} < \{1, 3, 5, 7\}$.

We'll refer to the index of a particular combination $\{r_1, r_2, \ldots, r_k\}$ of k elements, where $1 \leq r_i \leq n$ and $r_i < r_j$ for $i < j$, as $q(r_1, r_2, \ldots, r_k)$ or, when more convenient, $q(r_{12\ldots k})$.

Indexing combinations of n elements with $k = 1$

This is the trivial case. There are $\binom{n}{1} = n$ indices and it's obvious that

$$\boxed{q(r_1) = r_1\,, \qquad 1 \leq r_1 \leq n\,, \quad n \geq 1\,.}$$

Indexing combinations of n elements with $k = 2$

We want to find $q(r_1, r_2)$ for $1 \leq r_1 < r_2 \leq n$. We know that there should be $\binom{n}{2} = \frac{n(n-1)}{2}$ indices in total. We can represent all those combinations as elements in a matrix with elements $a_{ij} = (i,j)$, where $1 \le i \le n$ and $1 \le j \le n$, but with certain elements ignored, namely, those for which $i \ge j$:

$$\begin{pmatrix} - & (1,2) & (1,3) & \cdots & (1,n-2) & (1,n-1) & (1,n) \\ - & - & (2,3) & \cdots & (2,n-2) & (2,n-1) & (2,n) \\ - & - & - & \cdots & (3,n-2) & (3,n-1) & (3,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ - & - & - & \cdots & - & (n-2,n-1) & (n-2,n) \\ - & - & - & \cdots & - & - & (n-1,n) \\ - & - & - & \cdots & - & - & - \end{pmatrix}.$$

Their indices can then be collected in the corresponding matrix below,

$$\begin{pmatrix} - & 1 & 2 & \cdots & n-3 & n-2 & n-1 \\ - & - & n & \cdots & 2n-5 & 2n-4 & 2n-3 \\ - & - & - & \cdots & 3n-8 & 3n-7 & 3n-6 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ - & - & - & \cdots & - & \dots & \dots \\ - & - & - & \cdots & - & - & n(n-1)/2 \\ - & - & - & \cdots & - & - & - \end{pmatrix}.$$

It's clear that the last element in row r has index

$$q(r, c=n) = nr - \frac{r(r+1)}{2},$$

since the total number of missing elements up to and including row r is given by the sum of the first r integers, which itself equals $r(r+1)/2$. Moreover, the first non-ignored element of row r has index one higher than that of the last element of row $(r-1)$, so

$$q(r, c=r+1) = q(r-1, c=n) + 1 = n(r-1) - \frac{r(r-1)}{2} + 1$$

and we see that

$$q(r,c) = n(r-1) - \frac{r(r-1)}{2} + (c-r), \qquad 1 \le r < c \le n.$$

Note that the last index is

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

as expected. In the general notation we've adopted, the index of an arbitrary combination $\{r_1, r_2\}$ is, then,

$$\boxed{q(r_1,r_2) = n(r_1-1) - \frac{r_1(r_1-1)}{2} + (r_2-r_1), \qquad 1 \le r_1 < r_2 \le n, \quad n \ge 2.}$$

Indexing combinations of n elements with $k = 3$

We now have elements of the form (r_1, r_2, r_3) where $1 \le r_1 < r_2 < r_3 \le n$. Consider the case where $r_1 = 1$ so $2 \le r_2 < r_3 \le n$. This is similar to the previous case, with n replaced by $(n-1)$ so that $r_1 = 1$ comprises a total of $(n-1)(n-2)/2$ indices. This is more evident if we resort to our matrix representations where, now, the entire first row is ignored because neither r_2 nor r_3 can share the value of r_1 (1 in this instance):

$$\begin{pmatrix}
- & - & - & \cdots & - & - & - \\
- & - & (2,3) & \cdots & (2,n-2) & (2,n-1) & (2,n) \\
- & - & - & \cdots & (3,n-2) & (3,n-1) & (3,n) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
- & - & - & \cdots & - & (n-2,n-1) & (n-2,n) \\
- & - & - & \cdots & - & - & (n-1,n) \\
- & - & - & \cdots & - & - & -
\end{pmatrix}$$

and

$$\begin{pmatrix}
- & - & - & \cdots & - & - & - \\
- & - & 1 & \cdots & n-4 & n-3 & n-2 \\
- & - & - & \cdots & 2n-7 & 2n-6 & 2n-5 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
- & - & - & \cdots & - & \cdots & \cdots \\
- & - & - & \cdots & - & - & (n-1)(n-2)/2 \\
- & - & - & \cdots & - & - & -
\end{pmatrix}.$$

Therefore, when $r_1 = 1$, we find

$$q(r_1 = 1, r_2, r_3) = n(r_2 - 2) - \frac{r_2(r_2-1)}{2} + (r_3 - r_2) + 1\,, \qquad 1 < r_2 < r_3 \le n\,.$$

Consider next the case where $r_1 = 2$, so that $3 \le r_2 < r_3 \le n$. Now, the entire first *two* rows are ignored and the first non-ignored element is at row $r_2 = 3$ and column $r_3 = 4$:

$$\begin{pmatrix}
- & - & - & - & \cdots & - & - & - & - \\
- & - & - & - & \cdots & - & - & - & - \\
- & - & - & (3,4) & \cdots & (3,n-3) & (3,n-2) & (3,n-1) & (3,n) \\
- & - & - & - & \cdots & (4,n-3) & (4,n-2) & (4,n-1) & (4,n) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
- & - & - & - & \cdots & - & (n-3,n-2) & (n-3,n-1) & (n-3,n) \\
- & - & - & - & \cdots & - & - & (n-2,n-1) & (n-2,n) \\
- & - & - & - & \cdots & - & - & - & (n-1,n) \\
- & - & - & - & \cdots & - & - & - & -
\end{pmatrix}.$$

What's the index of element $(3,4)$? It's not 1 but, rather, 1 plus the total number of indices associated with the combinations having $r_1 = 1$, namely, $(n-1)(n-2)/2$. Taking that into account, we then find that the last element in row $r_2 > 2$ has index

$$q(r_1 = 2, r_2, r_3 = n) = \frac{(n-1)(n-2)}{2} + (r_2 - 2)n - \frac{r_2(r_2+1)}{2} + 3\,,$$

while the first non-ignored element of row r_2 has index

$$q(r_1 = 2, r_2, r_3 = r_2 + 1) = \frac{(n-1)(n-2)}{2} + (r_2 - 3)n - \frac{r_2(r_2-1)}{2} + 4\,,$$

resulting in

$$q(r_1 = 2, r_2, r_3) = \frac{(n-1)(n-2)}{2} + (r_2 - 3)n - \frac{r_2(r_2-1)}{2} + (r_3 - r_2) + 3\,, \quad 2 < r_2 < r_3 \leq n\,.$$

Next, let's take the case when $r_1 = 3$, so that $4 \leq r_2 < r_3 \leq n$. Now, the entire first *three* rows are ignored and the first non-ignored element is at row $r_2 = 4$ and column $r_3 = 5$. Once again, we have to take into account all the indices associated with elements having $r_1 = 1$ or $r_1 = 2$, and we have

$$q(r_1 = 3, r_2, r_3) = \frac{(n-1)(n-2)}{2} + \frac{(n-2)(n-3)}{2} + (r_2 - 4)n - \frac{r_2(r_2-1)}{2} + (r_3 - r_2) + 6\,,$$

where $3 < r_2 < r_3 \leq n$. A pattern is beginning to appear. For a given value of r_1, the first r_1 rows are entirely ignored and we have

$$q(r_1, r_2, r_3) = \sum_{s=1}^{r_1 - 1} \frac{(n-s)(n-s-1)}{2} + (r_2 - r_1 - 1)n - \frac{r_2(r_2-1)}{2} + (r_3 - r_2) + \frac{r_1(r_1+1)}{2}\,,$$

where $1 \leq r_1 < r_2 < r_3 \leq n$. Note that the total number of indices is

$$q(n-2, n-1, n) = \sum_{s=1}^{n-3} \frac{(n-s)(n-s-1)}{2} + 1 = \sum_{s=1}^{n-3} \binom{n-s}{2} + 1 = \sum_{m=2}^{n-1} \binom{m}{2}\,.$$

However, by the *hockey-stick identity*,

$$\sum_{m=k}^{n} \binom{m}{k} = \binom{n+1}{k+1}\,,$$

and we get

$$q(n-2, n-1, n) = \binom{n}{3}\,,$$

as we'd expect. Speaking of binomials, we can rewrite $q(r_1, r_2, r_3)$ like so:

$$\boxed{q(r_1, r_2, r_3) = \sum_{s=1}^{r_1-1} \binom{n-s}{2} + (r_2 - r_1 - 1)n - \frac{r_2(r_2-1)}{2} + (r_3 - r_2) + \frac{r_1(r_1+1)}{2}\,,}$$

where $1 \le r_1 < r_2 < r_3 \le n$ and $n \ge 3$.

Indexing combinations of n elements with $k = 4$

We now move on to $k = 4$ but the procedure is the same. We know that there should be a total of $\binom{n}{4}$ indices. For a given value of r_1, we need to account for all the indices associated with combinations that have their *first* element equal to 1, 2, and so on, up to but not including r_1. The total number of such indices are, respectively, $\binom{n-1}{3}$, $\binom{n-2}{3}$, $\binom{n-3}{3}$, and so on, up to but not including $\binom{n-r_1}{3}$. We then need to account for all the indices associated with combinations that have their *second* element equal to r_1+1, r_1+2, and so on, up to but not including r_2. The total number of such indices are, respectively, $\binom{n-r_1-1}{2}$, $\binom{n-r_1-2}{2}$, $\binom{n-r_1-3}{2}$, and so on, up to but not including $\binom{n-r_1-(r_2-r_1)}{2} = \binom{n-r_2}{2}$. Next, we need to account for all the indices associated with combinations that have their *third* element equal to r_2+1, r_2+2, and so on, up to but not including r_3. The total number of such indices are, respectively, $\binom{n-r_2-1}{1}$, $\binom{n-r_2-2}{1}$, $\binom{n-r_2-3}{1}$, and so on, up to but not including $\binom{n-r_2-(r_3-r_2)}{1} = \binom{n-r_3}{1}$. Finally, we must account for the value of r_4, which contributes the amount $(r_4 - r_3)$ to the index computation. These considerations result in

$$q(r_{1234}) = \sum_{s=1}^{r_1-1} \binom{n-s}{3} + \sum_{s=1}^{r_2-r_1-1} \binom{n-r_1-s}{2} + \sum_{s=1}^{r_3-r_2-1} \binom{n-r_2-s}{1} + (r_4 - r_3)\,.$$

Note that

$$(r_4 - r_3) = \sum_{s=1}^{r_4-r_3-1} \binom{n-r_3-s}{0} + 1$$

and, so,

$$\boxed{\begin{aligned} &q(r_{1234}) = 1 + \\ &\sum_{s=1}^{r_1-1} \binom{n-s}{3} + \sum_{s=1}^{r_2-r_1-1} \binom{n-r_1-s}{2} + \sum_{s=1}^{r_3-r_2-1} \binom{n-r_2-s}{1} + \sum_{s=1}^{r_4-r_3-1} \binom{n-r_3-s}{0}, \end{aligned}}$$

where $1 \le r_1 < r_2 < r_3 < r_4 \le n$ and $n \ge 4$.

Indexing combinations of n elements with an arbitrary k

The pattern is now clear and we can write a general expression for the index $q(r_{123\ldots k})$:

$$\boxed{q(r_1, r_2, \ldots, r_k) = 1 + \sum_{m=0}^{k-1} \sum_{s=1}^{\varphi(k,m)} \binom{n - r_{k-m-1} - s}{m},}$$

where $\varphi(k,m) \equiv r_{k-m} - r_{k-m-1} - 1$, $r_0 \equiv 0$, $1 \leq r_1 < r_2 < \cdots < r_k \leq n$, and $n \geq k \geq 1$.

Inverting the index

How can we obtain the values of $\{r_1, r_2, \ldots, r_k\}$ for a given value of $q(r_{123\ldots k})$? Well, we can't, not directly. What we *can* do, however, is perform a series of binary searches on the given index, since we know it's in sorted order. We know that the first element of an (n,k)-combination is an integer in the range $[1, n-k+1]$. If we compute the largest index with the first element equal to half that range and compare it with the given index, we'll be able to tell if the given index has its first element in the lower or in the upper half of that interval. Repeating this binary search will eventually zero-in on the value of the first element. Repeating this procedure for the remaining elements allows us to obtain the values of all the elements that gave rise to the given index. We'll need to perform k binary searches, each taking at most $\log_2(n)$ comparisons to pinpoint the element being searched for, for a total running time, in the worst case, of $k \log_2(n)$ comparisons.

For reference, here are expressions for the minimum and the maximum indices for which the first j elements are $\{r_1, r_2, \ldots, r_j\}$ $(1 \leq j \leq k)$:

$$\boxed{\begin{aligned} \min q(r_{123\ldots k} \,|\, r_{123\ldots j}) &\equiv q(r_1, r_2, \ldots, r_j, r_j + 1, r_j + 2, \ldots, r_j + k - j) \\ \max q(r_{123\ldots k} \,|\, r_{123\ldots j}) &\equiv q(r_1, r_2, \ldots, r_j, n - k + j + 1, n - k + j + 2, \ldots, n) \end{aligned}}$$

where the usual conditions apply on the free parameters. For example, for $n = 9$, $k = 5$, $r_1 = 2$, and $r_2 = 4$, the combination with the smallest index is $\{2, 4, 5, 6, 7\}$ while that with the maximum index is $\{2, 4, 7, 8, 9\}$.

■