Combinatorics: Information and Statistical Entropy

A problem that occurs often in statistical theory is a matter of counting the number Ω of different possibilities to select N_1 out of a total of N objects. This combinatorial problem is solved by the binomial coefficient

$$\Omega = \binom{N}{N_1} = \frac{N!}{N_1! \cdot (N - N_1)!} = \frac{N!}{N_1! \cdot N_2!} \tag{1}$$

Here, $N_2 = (N - N_1)$, and $N_1 + N_2 = N$ is the normalization. Because of the symmetry of expression (1), the quantity Ω is also the number of possibilities to select N_2 objects out of N total.

The problem relates to information theory, where one may ask the question, what information content can be transmitted by the two bits "0" and "1" in a train (word) of N total bits. Let the number of bits "0" be N_1 and N_2 that of bits "1". Then, Ω in Equ. 1 is the ***number of different words*** that can be transmitted with a length of N bits. If Ω is large, then the message contained in a given word can be very specific: ***The information is high for large Ω and low for small Ω.*** Hence, the task to maximize this quantity.

A similar task exists in biochemistry, where one is interested to evaluate the number of ways a few characteristic organic building blocks can be distributed over a DNA macro-molecule and what information can be relayed with these arrangements. In statistical theory, the number of ways a number N_1 of particles can be distributed over N states is of interest. The most probable distributions of particles are obtained for maximum entropy *S*, a variable defined as

$$S = k \cdot \ln \Omega \qquad (2)$$

Here, the quantity *k* is a universal constant (the Boltzmann constant $k = k_B$). This entropy can be maximized by numerical methods or, approximately for large numbers involved, also analytically. The case of statistical thermodynamics is particularly simple to evaluate, because the number of particles is large ($N_1 >> 1$) and the number of available states is even larger (often $N > N_1$). Then, one always can neglect numbers of the order of *1* in comparison with either N_1 or N_2. To evaluate Equ. 2, one may then use Stirling's formula,

$$\ln n! = n \cdot (\ln n - 1) \qquad (3)$$

Then, one can define the following measure of information

$$\begin{aligned} I &:= \ln \Omega = \ln N! - \ln N_1! - \ln N_2! = \\ &\approx N(\ln N - 1) - N_1 \ln(N_1 - 1) - N_2 \ln(N_2 - 1) \\ &\approx N \ln N - N_1 \ln N_1 - N_2 \ln N_2 \\ &= (N_1 + N_2) \ln N - N_1 \ln N_1 - N_2 \ln N_2 \end{aligned} \qquad (4)$$

or

$$I \approx -N_1 \ln \frac{N_1}{N} - N_2 \ln \frac{N_2}{N} \qquad (5)$$

This yields an information per word length *N* of

$$\frac{I}{N} \approx -\frac{N_1}{N} \ln \frac{N_1}{N} - \frac{N_2}{N} \ln \frac{N_2}{N} = -\sum_{n=1}^{2} p_n \ln p_n \qquad (6)$$

Here, the quantities

$$p_n = \frac{N_n}{N} < 1; \qquad (n = 1,2) \tag{7}$$

are the relative probabilities to find bit "0" ($n=1$) or bit "1" ($n=2$) in a word of length N.

If there are more than *2* objects to be distributed among N places, for example $M > 2$ objects, one generalizes Equ. 6 to

$$\frac{I}{N} = -\sum_{n=1}^{M} p_n \ln p_n; \;\; with \quad \sum_{n=1}^{M} p_n = 1 \tag{8}$$

To maximize the information I is equivalent to maximizing the information per bit (Equ. 8) by varying the probabilities p_n under the constraint of the normalization. This maximum is found in the usual way by setting all first derivatives

$$\frac{\partial}{\partial p_n}\left\{-\sum_{m=1}^{M} p_m \ln p_m - \lambda \left(\sum_{i=1}^{M} p_i - 1\right)\right\} = 0 \tag{9}$$

to zero, for $n=1,....M$. The normalization condition has been taken into account by adding it, multiplied by the ***Lagrange multiplier*** $\boldsymbol{\lambda}$, to the function to be maximized under this constraint. This procedure yields

$$-\ln p_n - p_n \cdot \frac{1}{p_n} - \lambda = 0 \tag{10}$$

or

$$\ln p_n = -(\lambda + 1) \tag{11}$$

and

$$p_n = e^{-(\lambda+1)} = p = const. \tag{12}$$

implying ***an equal probability for all n***. From the normalization condition, $\sum_{n=1,..M} p_n = 1$, follows immediately that

$$p_n = p = 1/M \tag{13}$$

This result implies that the maximum information can be transmitted in a word of length *N* bits, if all positions ***can*** carry a "0" or "1" with equal probability. This is expected, because if a bit position is "damaged", e.g., always carries bit "0", then it does not contribute to the information contained in the word.

In analogy to Equ. 8, the statistical entropy for a system of *N* particles distributed over *M* states is defined as

$$S = -k\sum_{n=1}^{M} p_n \ln p_n \geq 0 \tag{14}$$

where *k* is a universal constant not depending on the material properties of the system. The probabilities $p_n = N_n/N$ for states *n* to be populated fulfill the normalization condition

$$\sum_{n=1}^{M} p_n = \sum_{n=1}^{M} \frac{N_n}{N} = 1 \tag{15}$$

Following the same arguments as above for the information, one finds that the ***entropy is maximized for an equal population of all states with the probability $p_n = 1/M$***.