Entropy as the "Only Way to Measure Information"

Evgenii Samutichev

December 21, 2023

Theorem Let $H : (p_1, ..., p_n) \to H(p_1, p_2, ..., p_n)$ be a positive symmetric[a] function defined on finite probability distributions. If H has the following properties

(i) H is continuous in the p_i's

(ii) If all p_i areequal, $p_i = 1/n$, then H should be a monotonic increasing function of n (This means: "With equally likely events, there is more choice, or uncrertainty,when there are more possible events")

(iii)

$$H(p_1, p_2, ..., p_n) = H(p_1 + p_2, p_3, ..., p_n) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \tag{1}$$

which means: "If you first make a choice ignoring the difference between 1 and 2, and then, if 1 and 2 came up, you choose one of them, the original H should be the weighted sum of the individual values of H in the formula above"

Then, there exists a positive constant C s.t.

$$H(p_1, p_2, ..., p_n) = -C \sum_i p_i \log p_i \tag{2}$$

The choice of C amounts to the choice of a unit of measure.

[a] Reasonable additional assumption, without it I've no idea how to proceed

Proof. **Part 1** Property (1) can be extended by induction on the number of "glued together" values. If it's true for k, then for $k+1$ denote $p = p_1 + p_2 + ... + p_{k+1}$ and

$$\begin{aligned}
H(p_1, p_2, ..., p_k, p_{k+1}, ..., p_n) &= H(p_1 + p_2, p_3, ..., p_n) + (p_1 + p_2)H(\tfrac{p_1}{p_1+p_2}, \tfrac{p_2}{p_1+p_2}) = \\
&= H((p_1 + p_2) + ... + p_{k+1}, ..., p_n) + \\
&+ pH(\tfrac{p_1+p_2}{p}, \tfrac{p_3}{p}, ..., \tfrac{p_{k+1}}{p}) + (p_1 + p_2)H(\tfrac{p_1}{p_1+p_2}, \tfrac{p_2}{p_1+p_2}) = \\
&= H(p, ..., p_n) + pH(\tfrac{p_1}{p}, ..., \tfrac{p_{k+1}}{p}) - p\tfrac{p_1+p_2}{p}H(\tfrac{p_1}{p_1+p_2}, \tfrac{p_2}{p_1+p_2}) + \\
&+ (p_1 + p_2)H(\tfrac{p_1}{p_1+p_2}, \tfrac{p_2}{p_1+p_2}) = \\
&= H(p, ..., p_n) + pH(\tfrac{p_1}{p}, ..., \tfrac{p_{k+1}}{p})
\end{aligned}$$

For a positive integer n define $A(n) = H(\frac{1}{n}, ..., \frac{1}{n})$. Then

$$A(n^2) = H\left(\frac{1}{n^2}, ..., \frac{1}{n^2}\right) \tag{3}$$

but

$$\frac{1}{n} = \sum_{k=1}^{n} \frac{1}{n^2} \tag{4}$$

so we can use the extended property (iii) and symmetricity of H

$$\begin{aligned} A(n^2) = H(\tfrac{1}{n}, \tfrac{1}{n^2}, ..., \tfrac{1}{n^2}) + \tfrac{1}{n}H(\tfrac{1}{n}, ..., \tfrac{1}{n}) = ... = H(\tfrac{1}{n}, ..., \tfrac{1}{n}) + \\ + (\tfrac{1}{n} + ... + \tfrac{1}{n})H(\tfrac{1}{n}, ..., \tfrac{1}{n}) = A(n) + A(n) = 2A(n) \end{aligned}$$

Again, we continue by induction, if it's true for k that $A(n^k) = kA(n)$, then for $k+1$

$$A(n^{k+1}) = A(n^k) + (\underbrace{\tfrac{1}{n^k} + ... + \tfrac{1}{n^k}}_{n^k \text{ terms}})A(n) = kA(n) + A(n) = (k+1)A(n) \tag{5}$$

where we used

$$\frac{1}{n^k} = \sum_{k=1}^{n} \frac{1}{n^{k+1}} \tag{6}$$

and

$$\frac{\frac{1}{n^{k+1}}}{\sum_{k=1}^{n} \frac{1}{n^{k+1}}} = \frac{n^k}{n^{k+1}} = \frac{1}{n} \tag{7}$$

An immediate consequence of that property is $A(1) = 0$, since

$$A(1) = A(1^2) = 2A(1) \implies A(1) = 0 \tag{8}$$

Part 2

Intervals

$$[0, \log m], [\log m, 2\log m], ..., [\ell \log m, (\ell+1)\log m], ... \tag{9}$$

cover the entire nonnegative real line. Thus, for $k \log n \geq 0$ there exists ℓ s.t.

$$\ell \log m \leq k \log n \leq (\ell + 1)\ ogm \tag{10}$$

and by exponentiating we get that for all $m > 1, n > 1$ and $k \geq 1$ there exists $\ell \in \mathbb{N}$ s.t.

$$m^\ell \leq n^k \leq m^{\ell+1} \tag{11}$$

Now, using the property we have proven in **part 1** and assumption (ii) giving us that $A(n)$ is monotonic increasing function of n, we get

$$A(m^\ell) \leq A(n^k) \leq A(m^{\ell+1}) \tag{12}$$

and

$$\ell A(m) \leq kA(n) \leq (\ell+1)A(m) \tag{13}$$

Dividing both sides by $A(m) > A(1) = 0$ and $k \geq 1$ we derive

$$\frac{\ell}{k} \leq \frac{A(n)}{A(m)} \leq \frac{\ell+1}{k} \tag{14}$$

From (10) we can similarly derive

$$\frac{\ell}{k} \leq \frac{\log n}{\log m} \leq \frac{\ell+1}{k} \tag{15}$$

And combining these two inequalities together, we conclude that

$$\left| \frac{A(n)}{A(m)} - \frac{\log n}{\log m} \right| \leq \frac{1}{k} \tag{16}$$

notice how it doesn't depend on ℓ anymore and k is arbitrary.

Thus, by letting $k \to \infty$ we have

$$\frac{A(n)}{A(m)} = \frac{\log n}{\log m} \tag{17}$$

Rearranging terms we get that for all $n, m > 1$

$$\frac{A(n)}{\log n} = \frac{A(m)}{\log m} = C \tag{18}$$

where $C > 0$ is some constant. So

$$A(n) = C \log n \tag{19}$$

and this is true for all $n \geq 1$, since for $n = 1$ it's trivially true.

Part 3 We can apply the extended property (iii) of H, that we proved in **part 1**. If $p_i = n_i / \sum_{j=1}^{n} n_j$ for some integers $n_1, ..., n_n$, then defining $p = \sum_{j=1}^{n_j}$ gives us

$$A(p) = H\left(\frac{1}{p}, ..., \frac{1}{p}\right) = H\left(\sum_{j=1}^{n_1} \frac{1}{p}, ..., \sum_{j=1}^{n_n} \frac{1}{p}\right) + \sum_{i=1}^{n} H\left(\frac{1}{n_i}, ..., \frac{1}{n_i}\right) \sum_{j=1}^{n_i} \frac{1}{p} \tag{20}$$

and since by definition

$$A(n_i) = H\left(\frac{1}{n_i}, ..., \frac{1}{n_i}\right) \tag{21}$$

and also

$$\sum_{j=1}^{n_i} \frac{1}{p} = \frac{n_i}{p} = p_i \tag{22}$$

we conclude by applying the result of **part 2** that

$$C \log \sum_{j=1}^{n} p_i = C \log p = H(p_1, ..., p_n) + C \sum_{i=1}^{n} p_i \log n_i \tag{23}$$

We can rewrite the left hand side as

$$C \sum_{i=1}^{n} p_i \log \sum_{j=1}^{n} p_i \tag{24}$$

since $\sum_{i=1}^{n} p_i = 1$. Then, by moving the second term of the r.h.s to the left we get

$$H(p_1, ..., p_n) = -C \sum_{i=1}^{n} p_i \log \frac{n_i}{p} = -C \sum_{i=1}^{n} p_i \log p_i \tag{25}$$

Part 4 In fact, we have just proven the result for all discrete probability distributions involving PMF with rational values. That is because we can always make them to have the common denominator. If

$$p_i = \frac{k_i}{m_i}, \quad i = 1, ..., n \tag{26}$$

s.t.

$$\sum_{i=1}^{n} p_i = 1 \tag{27}$$

then

$$p_i = \frac{k_i}{m_i} = \frac{k_i \prod_{j=1, j\neq i}^{n} m_j}{m_1 ... m_n} \tag{28}$$

and by normalization

$$\sum_{i=1} n k_i \prod_{j=1, j\neq i}^{n} m_j = m_1 ... m_n \tag{29}$$

so we can take

$$n_i = k_i \prod_{j=1, j\neq i}^{n} m_j \tag{30}$$

reducing to **part 3**. As we know, rational numbers are dense in real numbers, so we can simply approximate any sequence $p_1, ..., p_n$ s.t. $\sum_{i=1}^{n} p_i = 1$ by n sequences of rational numbers, while also normalizing them. That is, we have $\{r_j^{(i)}\}_{j=1}^{\infty}$ for each $i = 1, ..., n$ s.t. $r_j^{(i)} \to p_i$ as $j \to \infty$. But if we normalize, then

$$h_j^{(i)} = \frac{r_j^{(i)}}{\sum_{k=1}^{n} r_j^{(k)}} \to \frac{p_i}{1} = p_i \tag{31}$$

and by continuity assumption (i) of H we then have

$$H(h_j^{(1)}, ..., h_j^{(n)}) \to H(p_1, ..., p_n) \tag{32}$$

while at the same time

$$C \sum_{i=1}^{n} h_j^{(i)} \log h_j^{(i)} \to C \sum_{i=1}^{n} p_i \log p_i \tag{33}$$

so from (25) we get by taking the limit, that for all discrete finite probability distributions

$$H(p_1, ..., p_n) = -C \sum_{i=1}^{n} p_i \log p_i \tag{34}$$

$\square$