Exercise: We derive the Shannon entropy as a measure of the uncertainty represented by a discrete probability distribution.

(Source: Jaynes, Information Theory and Statistical Mechanics)

Suppose we have a variable x which can take values $\{x_1, \dots, x_n\}$. Our incomplete understanding of the processes which determine the value of x can be represented by assigning probability p_i to outcome x_i, for $i \in \{1, \dots, N\}$

We wish to find a function H to characterize the "uncertainty" represented by the probability distribution $\{p_i\}_{i=1}^{n}$. We require H to have 3 properties

1.) $H(p_1, \dots, p_n)$ is continuous

2.) When all outcomes are equally probable (i.e. $p_i = 1/n \ \forall i \in \{1, \dots, n\}$),

$$H(1/n, \dots, 1/n) \equiv A(n)$$ is a monotonically increasing function of n.

(Intuitively, a uniform distribution over more outcomes is more "uncertain" than a uniform distribution over fewer values.)

3.) Composition Law

Suppose we wish to group outcomes together. We might group $\{x_1, \dots, x_k\}$ together with probability $W_1 = p_1 + \dots + p_k$, $\{x_{k+1}, \dots, x_{k+m}\}$ with probability $W_2 = p_{k+1} + \dots + p_{k+m}$, and so on. We can then assign conditional probabilities as $p(x_1|W_1) = p_1/W_1, \dots, p(x_k|W_1) = p_k/W_1$, and so on. Specifying the group probabilities and the conditional probabilities is equivalent to specifying the original probabilities, so we require the uncertainty to be the same in both cases.

$$H(p_1, \dots, p_n) = H(W_1, \dots, W_r) + W_1 H(p_1/W_1, \dots, p_k/W_1)$$

$$+ W_2 H(p_{k+1}/W_2, \dots, p_{k+m}/W_2) + \dots$$

We show that the Shannon entropy satisfies these requirements.

Proof|

Suppose we have probabilities $\{p_i\}_{i=1}^n$. By (1), we only need to consider rational probabilities, as irrational numbers can be constructed from sequences of rational numbers, and H is continuous, so it will behave well under such sequence limits. Thus, we can find numbers $\{n_i\}_{i=1}^{N}$, $n_i \in \mathbb{N}$ such that $p_i = n_i / \sum_i n_i$.

We will now treat these probabilities as "groups" made from some uniform distribution over $\sum_i n_i$ numbers. By rule (2), this uniform distribution has uncertainty $A(\sum_i n_i)$.

By rule (3), we have

$$A\Big(\sum_i n_i\Big) = H(p_1, \dots, p_n) +$$

$$\sum_j p_i \underbrace{H\left(\frac{1/\sum_i n_i}{n_j/\sum_i n_i}, \dots, \frac{1/\sum_i n_i}{n_j/\sum_i n_i}\right)}_{A(n_j)} \Rightarrow$$

$$H(p_1, \dots, p_n) = A\Big(\sum_i n_i\Big) - \sum_i p_i A(n_i).$$

To determine A, we consider the case when $n_i = m \ \forall i \in \{1, \dots, n\}$.

Then, $\sum_i n_i = m \cdot n$, and

$$H\left(\frac{m}{mn}, \dots, \frac{m}{mn}\right) = A(mn) - \sum_i \frac{m}{mn} A(m)$$

$$\iff A(n) + A(m) = A(m \cdot n)$$

We thus conclude that $A(n) = K \log(n)$, where $K > 0$ by the (2). Finally, substituting into the formula for $H(p_1, \dots, p_n)$, we derive

$$H(p_1, \dots, p_n) = K \log\left(\sum_i n_i\right) - K \sum_i p_i \log(n_i) =$$

$$K \log\left(\sum_i n_i\right) - K \sum_i \left[p_i \log(n_i) - p_i \log\left(\sum_i n_i\right) + p_i \log\left(\sum_i n_i\right)\right]$$

$$= K \log\left(\sum_i n_i\right) - K \sum_i p_i \log\left(n_i / \sum_i n_i\right) - K \log\left(\sum_i n_i\right) \sum_i p_i$$

$$= -K \sum_i p_i \log p_i \iff$$

$$\boxed{H(p_1, \dots, p_n) = -K \sum_{i=1}^{n} p_i \log p_i}$$