

# Measuring Randomness

Sarang Sane

*Based on a talk in the students' seminar*

---

## Abstract

Consider throwing a "fair" coin as opposed to throwing a coin with two heads. It is natural to think that the experiment/ action/event of throwing a "fair" coin is "more random" than that of throwing the coin with two heads. This takes us to the natural question of whether we can somehow measure randomness.

For the major portion of this talk, we will try to answer the question, "How random is this action/experiment/phenomenon"? This part is intended to be elementary, precise and completely self-contained. The answer will be an expression which occurs in many other contexts as well. Towards the end, I will try and make an ambitious attempt to try and tie this up with applications in science (this part might be very vague).

---

This write-up is based on a student seminar I gave in which we studied Shannon entropy as a measure of randomness (we formulate axioms that such a function should follow based on heuristics and then prove Shannon's result that there is a unique such function) and then looked at one of the contexts in which this problem arose (coding theory).

Let  $X$  denote our experiment which can take  $n$  values  $x_1, x_2, \dots, x_n$  with probabilities  $p_1, p_2, \dots, p_n$ . Note that for our purpose, the data  $p_1, p_2, \dots, p_n$  is enough to come up with a measure of randomness (in other words, the values  $x_1, x_2, \dots, x_n$  have no bearing on how random the experiment is). Let  $H$  be our measure of the randomness in the experiment. Since the values  $x_1, x_2, \dots, x_n$  have

no role,  $H$  is a function of  $p_1, p_2, \dots, p_n$  (denoted by  $H(p_1, p_2, \dots, p_n)$  or  $H(X)$  as per convenience).

Then, we make the following heuristic demands of  $H(p_1, p_2, \dots, p_n)$  :

1.  $H$  is symmetric in  $p_1, p_2, \dots, p_n$
2.  $H(p_1, p_2, \dots, p_n) \geq 0$
3.  $H(p_1, p_2, \dots, p_n) = 0$  if and only if  $p_i = 1$  for some  $i$
4.  $H(p_1, p_2, \dots, p_n)$  achieves its maximum at  $p_1 = p_2 = \dots = p_n = \frac{1}{n}$
5.  $H(p_1, p_2, \dots, p_n) = H(p_1, p_2, \dots, p_n, 0)$
6.  $H(p_i q_j : i = 1, 2, \dots, m; j = 1, 2, \dots, n) = H(p_1, p_2, \dots, p_m) + H(q_1, q_2, \dots, q_n)$   
(Randomness of independent experiments equals the sum of the individuals)

Let us now see what we obtain from these 6 axioms. Let

$$h(n) = H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)$$

From (6),

$$\begin{aligned} h(n^r) &= H\left(\frac{1}{n^r}, \frac{1}{n^r}, \dots, \frac{1}{n^r}\right) \\ &= rH\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) \\ &= rh(n) \end{aligned}$$

Then,  $h(n)$  is monotonic since

$$\begin{aligned} h(n) &= H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) \\ &= H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}, 0\right) \\ &\leq H\left(\frac{1}{n+1}, \frac{1}{n+1}, \dots, \frac{1}{n+1}\right) \\ &= h(n+1) \end{aligned}$$

Let  $k, l \in \mathbb{N}$  such that  $k, l > 1$ . Then, given  $\epsilon > 0$ , choose  $r$  such that  $1/r < \epsilon$ , where  $r \in \mathbb{N}$ . Then,  $\exists s \in \mathbb{N}$  such that

$$\begin{aligned} l^s &\leq k^r \leq l^{s+1}. \\ \Rightarrow h(l^s) &\leq h(k^r) \leq h(l^{s+1}). \\ \Rightarrow sh(l) &\leq rh(k) \leq (s+1)h(l). \\ \Rightarrow \frac{s}{r} &\leq \frac{h(k)}{h(l)} \leq \frac{s+1}{r} \end{aligned}$$

Since logarithm satisfies these inequalities as well,

$$\frac{s}{r} \leq \frac{\log k}{\log l} \leq \frac{s+1}{r}$$

Hence,

$$\left| \frac{\log k}{\log l} - \frac{h(k)}{h(l)} \right| \leq \frac{1}{r} < \epsilon$$

Since  $\epsilon$  was arbitrary,  $\frac{\log k}{\log l} = \frac{h(k)}{h(l)}$ . Thus,  $\frac{h(k)}{\log k}$  is a constant !!! Hence, if we make a choice of unit,  $h(n)$  will be completely determined. The natural choice is to let the fair coin toss be our unit, i.e.

$$7. \quad h(2) = 1.$$

Then  $h(k) = \log_2 k$ . Thus, from the hypotheses(1)-(7), we obtain

$$H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = \log_2 n$$

Notice that we have not said anything about what we expect when we have experiments which are not independent. For those of us familiar with conditional probability (and actually for even those who are not), the following heuristics will not be hard to digest. Namely, we measure the randomness of one of the experiments and then measure the randomness of the other **CONDITIONAL** on the earlier one having already occurred. This means that conditional on each possible outcome of the first experiment, we calculate the randomness of the second and then sum these up with weights as the probabilities of the outcomes of the first. Putting this into equations, we get the following :

$$H((p_{ij})) = H(p_{1*}, p_{2*}, \dots, p_{n*}) + \sum_{i=1}^n p_{i*} H\left(\frac{p_{i1}}{p_{i*}}, \frac{p_{i2}}{p_{i*}}, \dots, \frac{p_{im}}{p_{i*}}\right)$$

where  $p_{i*} = \sum_{j=1}^n p_{ij}$ .

As self-respecting mathematicians, we of course that  $H$  be continuous! This leads us to the following :

**Theorem 1** (C.Shannon, 1948). *Let  $H$  be a function of  $(p_1, p_2, \dots, p_n)$ ;  $\sum p_i = 1, n \in \mathbb{N}$  satisfying the following properties :*

1.  $H$  is symmetric.
2.  $H \geq 0$
3.  $H(p_1, p_2, \dots, p_n) = 0$  if and only if  $p_i = 1$  for some  $i$
4.  $H(p_1, p_2, \dots, p_n)$  achieves its maximum at  $p_1 = p_2 = \dots = p_n = \frac{1}{n}$
5.  $H(p_1, p_2, \dots, p_n)$  is continuous in  $(p_1, p_2, \dots, p_n)$
6.  $H((p_{ij})) = H(p_{1*}, p_{2*}, \dots, p_{m*}) + \sum_{i=1}^m p_{i*} H\left(\frac{p_{i1}}{p_{i*}}, \frac{p_{i2}}{p_{i*}}, \dots, \frac{p_{in}}{p_{i*}}\right)$
7.  $H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$

Then  $H(p_1, p_2, \dots, p_n) = -\sum_{i=1}^n p_i \log_2 p_i$  is the unique function satisfying these properties.

*Proof.* Observe first that if  $p_{ij} = p_{i*} p_{*j}$ , where  $p_{i*} = \sum_{j=1}^n p_{ij}$ ,  $p_{*j} = \sum_{i=1}^m p_{ij}$ , (i.e., if the experiments are independent), then (6) gives us that

$$H((p_{ij})) = H(p_{1*}, p_{2*}, \dots, p_{m*}) + H(p_{*1}, p_{*2}, \dots, p_{*n}).$$

This was axiom (6) of the original set of axioms (refer page 2). Note that all the other axioms there and the choice of unit ( $H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$ ) are part of the hypotheses. Hence, we already know that

$$H(1/n, 1/n, \dots, 1/n) = \log_2 n.$$

Now consider an experiment  $(p_1, p_2, \dots, p_n)$  where all  $p_i$  are rational. Then  $p_i = m_i/m$  for some  $m_i, m \in \mathbb{N}$ . Let  $m_0 = 0$ . Define a new experiment  $Y$  which takes values in the set  $\{1, 2, \dots, m\}$ . We define the probabilities conditional on  $X = x_i$  as uniform on  $m_1 + m_2 + \dots + m_{i-1} + 1, m_1 + m_2 + \dots + m_{i-1} + 2, \dots, m_1 + m_2 + \dots + m_{i-1} + m_i$ , i.e.,

$$P(Y = j/X = x_i) = \begin{cases} 1/m_i & \text{if } j \in [\sum_{r=0}^{i-1} m_r + 1, \sum_{r=0}^i m_r] \cap \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

Hence,

$$p_{ij} = P(X = x_i, Y = j) = \begin{cases} p_i \times \frac{1}{m_i} = \frac{1}{m} & \text{if } j \in [\sum_{r=0}^{i-1} m_r + 1, \sum_{r=0}^i m_r] \cap \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

Hence, for a fixed  $j$ , there is a unique  $i_0$  for which  $p_{i_0 j} = 1/m$  and for all  $i \neq i_0$ ,  $p_{ij} = 0$ . This shows that

$$p_{*j} = 1/m \quad \forall j \in 1, 2, \dots, m.$$

Hence,

$$\begin{aligned} H((p_{ij})) &= H(p_{*1}, p_{*2}, \dots, p_{*m}) + \sum_{j=1}^m p_{*j} H\left(\frac{p_{1j}}{p_{*j}}, \frac{p_{2j}}{p_{*j}}, \dots, \frac{p_{mj}}{p_{*j}}\right) \\ &= H\left(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}\right) + \sum_{j=1}^m \frac{1}{m} H(0, \dots, 0, 1, 0, \dots, 0) \\ &= \log_2 m + 0 \\ &= \log_2 m \end{aligned}$$

But computing the same quantity in the other way, we get :

$$\begin{aligned}
H((p_{ij})) &= H(p_{1*}, p_{2*}, \dots, p_{n*}) + \sum_{i=1}^n p_{i*} H\left(\frac{p_{i1}}{p_{i*}}, \frac{p_{i2}}{p_{i*}}, \dots, \frac{p_{im}}{p_{i*}}\right) \\
&= H(p_1, p_2, \dots, p_n) + \sum_{i=1}^n \frac{m_i}{m} H\left(0, \dots, 0, \frac{1}{m_i}, \dots, \frac{1}{m_i}, 0, \dots, 0\right) \\
&= H(p_1, p_2, \dots, p_n) + \sum_{i=1}^n \frac{m_i}{m} H\left(\frac{1}{m_i}, \dots, \frac{1}{m_i}\right) \\
&= H(p_1, p_2, \dots, p_n) + \sum_{i=1}^n \frac{m_i}{m} \log_2 m_i
\end{aligned}$$

Putting both calculations together, we get that

$$\begin{aligned}
H(p_1, p_2, \dots, p_n) &= \log_2 n - \sum_{i=1}^n \frac{m_i}{m} \log_2 m_i \\
&= \sum_{i=1}^n \frac{m_i}{m} \log_2 \frac{m_i}{m} \\
&= - \sum_{i=1}^n p_i \log_2 p_i
\end{aligned}$$

exactly as we wanted. Recall that we did this under the assumption that  $p_i \in \mathbb{Q} \forall i$ . But  $H$  was assumed to be continuous and the right hand side is also continuous. Since the rationals are dense, we finally obtain

$$H(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \log_2 p_i.$$

□

**The function  $H$  is called the (Shannon) entropy.**

Shannon introduced the concept of entropy to study the theory of languages as well. Consider a set of alphabets, that is, any non-empty set. Then we look at the word semi-group formed under composition or joining. We call this a language. Suppose I have some messages in this language, say  $m$  many messages in language generated by alphabet  $A$ . I wish to code these in some way, say in the language

$B$  which has  $D$  many characters. Let  $C_j$  be the codeword corresponding to the message  $j$ , with length  $l_j$ . Now, we can calculate the frequency of occurrence of a particular alphabet or word or set of words in a language and based on that, assign a probability measure on the set of words in the language. For example, in the english language it is known that some letters are more frequent than others (for example, a,t,e are most frequent and z,x,q are least frequent). Suppose we have such a probability measure  $\underline{p}$  on the  $m$  messages and  $\underline{p} = (p_1, p_2, \dots, p_m)$ . Thus, we have

<i>Probabilities</i>	<i>Messages</i>		<i>Codeword</i>	<i>CodewordLengths</i>
$p_1$	1		$C_1$	$l_1$
$p_2$	2		$C_2$	$l_2$
$\vdots$	$\vdots$	$\xrightarrow{C}$	$\vdots$	$\vdots$
$p_m$	$m$		$C_m$	$l_m$

The basic objective of all this is that given the languages  $A$  and  $B$ , and the probabilities, we try to get a code  $C$  which has the minimum average length, that is, we try to get a code minimizing

$$\bar{l}(C) = \sum_{i=1}^m p_i l_i.$$

A question which would still be gnawing at the back of one's mind would be where entropy comes into the picture in this entire discussion. Shannon proved that for any  $C$ ,

$$\bar{l}(C) \geq - \sum_{i=1}^m p_i \log_2 p_i.$$

He further proved the existence of a code  $C_0$  with  $\bar{l}(C_0) \leq - \sum_{i=1}^m p_i \log_2 p_i + 1$ .

Now note that the set of all words form a semi-group under joining and that the null word is included to serve as identity. Then the code map  $C$  is called **irreducible** if no codeword  $C_j$  is an extension of a codeword  $C_i, i \neq j$ . It is called

**uniquely decipherable** if the extension of  $C$  to the word semi-group over  $\{1, 2, \dots, m\}$  is 1-1. Now, clearly any irreducible code is uniquely decipherable. Let the words  $i_1 i_2 \dots i_k$  and  $j_1 j_2 \dots j_l$ , where each  $i_x$  and  $j_y \in \{1, 2, \dots, m\}$  have the same codeword. Then, consider  $i_1$  and  $j_1$ . They must have the same codeword and hence must be same, else it will be a contradiction to irreducibility. In this way, one can continue for 2,3, and so on. This will prove that  $k = l$  and  $i_x = j_x$  and hence the extension of  $C$  to the word semigroup exists and is 1-1. Hence, the code  $C$  is uniquely decipherable. However, the converse is not true. To see this, choose  $m = 2$  and  $B = \{0, 1\}$ . Let  $C(1) = 0$  and  $C(2) = 01$ . Then,  $01$  is an extension of  $0$  and hence this code is not irreducible. However, observe that it is uniquely decipherable.

Here is a theorem which gives an equation to help decipher when irreducible codes will actually exist.

**Theorem 2.** (*Kraft's Inequality*) *An irreducible code, with lengths  $l_1, l_2, \dots, l_m$  exists if and only if*

$$D^{-l_1} + D^{-l_2} + \dots + D^{-l_m} \leq 1. \quad \dots (*)$$

Now, suppose that  $D = 2$ . This is a situation we are familiar with...where our messages are being coded as strings of 0 and 1. In this situation, there is an algorithm which allows to actually get an irreducible code which has minimum code length!!! We describe this algorithm below and give a few illustrative examples.

W.l.g. assume that  $p_1 \geq p_2 \geq \dots \geq p_m > 0$ . Suppose that  $l_1, l_2, \dots, l_m$  achieves minimum length. Then, being irreducible,

$$2^{-l_1} + 2^{-l_2} + \dots + 2^{-l_m} \leq 1.$$

Suppose for some  $i < j, l_i > l_j$ . Then, interchanging the codewords for the  $i$ -th and  $j$ -th letters, we get a new irreducible code (since the above equation is necessary and sufficient) with average length strictly lesser than the original one. But the original code was assumed to be optimal and this gives a contradiction. Hence, the optimal code must be such that  $l_1 \leq l_2 \leq \dots \leq l_m$ . Further, suppose that  $l_m > l_{m-1}$ . Then, the codeword corresponding to the  $m$ -th message say  $w_m$  is

of the form  $c_1c_2 \dots c_{l_m}$  where each  $c_i \in \{0, 1\}$  and can be written as  $c_1c_2 \dots c_{l_{m-1}}w'$ . But then we can simply drop  $w'$  and the resulting codeword will be shorter, while the code still remains irreducible which will mean the new code has a strictly lesser average length, once again contradicting optimality of the original code. Hence,  $l_m = l_{m-1}$ . Thus, the optimal irreducible code should satisfy  $l_1 \leq l_2 \leq \dots \leq l_{m-1} = l_m$ . Also, by irreducibility, it also follows that

$$2^{-l_1} + 2^{-l_2} + \dots + 2^{-l_{m-2}} + 2^{-(l_{m-1}-1)} \leq 1.$$

Now, suppose we club the last two messages together and form a new message  $m-1, m$ . Clearly, the new set-up is,

$$\begin{array}{cccccc} 1 & 2 & \dots & m-2 & \{m-1, m\} \\ p_1 & p_2 & \dots & p_{m-2} & p_{m-1} + p_m \end{array}$$

and the previous inequality is satisfied. Hence,  $\exists$  an irreducible code  $C'$  for the reduced set of messages such that length of the  $i$ -th codeword is  $l_i$  for  $i$  from 1 to  $m-2$  and for the combined final message the codeword length is  $l_{m-1} - 1$ . Let  $C_0'$  be the optimal code for the combined messages with codewords  $w_1', w_2', \dots, w_{m-2}', w_{m-1}'$  and corresponding lengths of the codewords  $k_1, k_2, \dots, k_{m-1}$ . Make a new code  $E$  for the original messages  $w_1', w_2', \dots, w_{m-2}', w_{m-1}'0, w_{m-1}'1$ . Since  $C_0'$  is irreducible, so is  $E$ . Hence, we get the two equations  $\bar{l}(C') \leq \bar{l}(C_0')$  and  $\bar{l}(C) \leq \bar{l}(E)$  which means,

$$p_1k_1 + p_2k_2 + \dots + p_{m-2}k_{m-2} + (p_{m-1} + p_m)k_{m-1} \leq$$

$$p_1l_1 + p_2l_2 + \dots + p_{m-2}l_{m-2} + (p_{m-1} + p_m)(l_{m-1} - 1)$$

and

$$p_1k_1 + p_2k_2 + \dots + p_{m-2}k_{m-2} + (p_{m-1} + p_m)(k_{m-1} + 1) \geq$$

$$p_1l_1 + p_2l_2 + \dots + p_{m-2}l_{m-2} + p_{m-1}l_{m-1} + p_ml_m.$$

Now, using both the equations with the fact that  $l_m = l_{m-1}$ , we get that there is equality in both equations. But this means that  $E$  is also an optimal code!!!

**Thus, to construct the optimal code for  $m$  messages with probability distribution  $p_1 \geq p_2 \geq \dots \geq p_m$ , we first construct it for  $m-1$  messages**

**with distribution  $p_1, p_2, \dots, p_{m-1}+p_m$  and if that is  $w_1', w_2', \dots, w_{m-2}', w_{m-1}'$ , then the optimal code for the original  $m$  messages is**

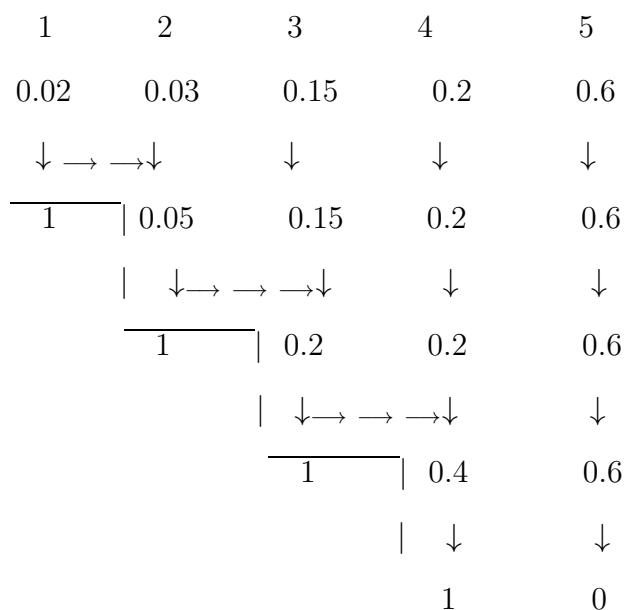
$w_1', w_2', \dots, w_{m-2}', w_{m-1}'0, w_{m-1}'1$ .

One can now proceed by induction. Note that one has to be a bit careful because one has to ensure that at each stage the condition  $p_1 \geq p_2 \geq \dots \geq p_m > 0$  is satisfied. If not, then one has to relabel the terms so that the condition is satisfied. This entire process is exactly Huffman's algorithm.

Once again, the key sentence to keep in mind is, at each stage, **JOIN VER-TICES WITH MINIMUM PROBABILITIES** .

Just to illustrate it, let us consider a few examples. (The pictures below are not very good. It is better to view these as binary trees. I will try to modify these with better ones in the near future.)

1.

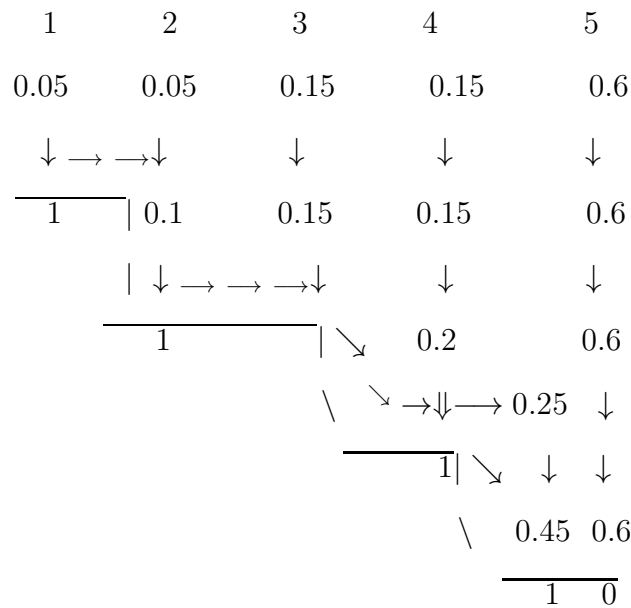


Now retrace the arrows to the required digit, and take the digits below the arrows in the same order. That will give you the codeword for the digit. Then, the Huffman code for this set-up is,

$$\begin{array}{lll}
5 \rightarrow 0, & 4 \rightarrow 10, & 3 \rightarrow 110, \\
2 \rightarrow 1110, & 1 \rightarrow 1111 &
\end{array}$$

Note here that the probabilities were ordered throughout the procedure. We consider an example where that condition breaks down mid-way. We interchange the digits (messages) to preserve that condition and carry on the procedure.

2.



This time while retracing note that we have interchanged at one point, (where it is represented by the double arrow) and so one should be careful in tracing back the arrows. The Huffman code for this set-up thus is,

$$\begin{array}{l}
5 \rightarrow 0 \\
4 \rightarrow 11 \\
3 \rightarrow 100 \\
2 \rightarrow 1010 \\
1 \rightarrow 1011
\end{array}$$

*Good Reference : Ergodic Theory and Information by Patrick Billingsley.*