

Appendix 1: Shannon Entropy and Information

In chapter 1, Shannon entropy is mentioned in passing. Shannon entropy is intimately tied to information theory. For readers who may read more about information theory elsewhere, this section will prove useful.

Information is transmitted in symbols. These symbols are most frequently letters. The English alphabet is a group of symbols. These symbols can be arranged to contain information. The average information per symbol is called Shannon entropy. It is almost always represented in the literature by a capital H.

H = Shannon Entropy

If all symbols are equally probable, the Shannon entropy is defined as:

$$H = 3.32 \times \log(\text{number of possible symbols}) \quad \text{Eq. 1}$$

or for readers with a base 2 log function on their calculator

$$H = \log_2(\text{number of possible symbols})$$

For example, if 26 blocks are labeled with letters from the alphabet and then drawn from a hat, the average information per symbol is $H = 3.32 \times \log(26) = 4.7$ bits per letter.

Now suppose that 102 blocks with the letter Z are added to the hat in the previous example. The hat now contains 128 blocks, but most are labeled with a Z. The Shannon entropy is now the weighted average of the contents in the hat. The odds of pulling the letter A from the hat are 1 in 128. So from equation 2, in chapter 1, the information associated with drawing an A is $3.32 \times \log(128/1) = 7$ bits. With the exception of the letter Z, all other letters have the same odds. Thus, observing any letter, A-Y, results in 7 bits of information.

The letter Z will be drawn from the hat 102 times with every 128 tries. This corresponds to 1 time in 1.25 tries, and the corresponding information is $3.32 \times \log (128/102) = 1.6$ bits.

Shannon entropy is the weighted average. 26 symbols contribute 7 bits and 102 contribute 1.6 bits. So H is calculated as follows: $(102/128) \times 1.6 + 26/128 \times (7) = 1.275 + 1.42 = 2.7$ bits per symbol. This means that a code exists that can transmit the contents of the hat using on average only 2.7 bits per symbol.

Question: How much information is carried by this message: “AAAAAAAA” if it is drawn from the hat with 128 letters?

Answer: Each A contributes 7 bits, so this message contains 8 letters x 7 bits per letter = 56 bits of information.

Question: what are the odds of drawing it from the hat?

Answer: 1 in 2^{56} or 1 in 7.2×10^{16} .

Question: on average how much information will most 80 letter messages drawn from this hat contain?

Answer: 80 letters x 2.7 bits per letter = 216 bits.

Question: How much information is carried by this message: “ZZZZZZZZ”?

Answer: Each Z contributes 1.6 bits. So this message contains 8 letters x 1.6 bits per letter = 12.8 bits of information.

Shannon entropy only works for very long messages. Short messages may or may not be accurately represented by Shannon entropy. The information content of a short message may be much higher or much lower. The examples above illustrate this concept. Long messages will always converge to the average information per bit. This is why Shannon entropy is useful.

<http://www.theory-of-evolution.net>