Solutions 6: Entropy

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Definition, interpretation and usual distributions

a) See Jupyter notebook. To compute the entropy of X, we only need the probability distribution of X. For instance the possible values x of X do not enter, only the probabilities $P(X = x)$ do.

b) In the case of a Bernoulli random variable with parameter $p = 1$, one probability value is 0. This case must be handled separately because $\log_2(0)$ is not defined (since $\lim_{x\to 0} \log_2(x) = -\infty$). If it is not, Python gives an error message and the result is `nan`. The contribution to entropy of a state x such that $P(x) = 0$ should be 0. Indeed, formally, $\lim_{x\to 0} x \log_2(x) = 0$, and qualitatively a state never observed should not increase the entropy of a random variable.

c) See Jupyter notebook.

d) Consider a random variable X with two outcomes, 0 and 1, such that $P(0) = p$ and $P(1) = 1 - p$. Its entropy is

$$H(X) = -p \log_2(p) - (1 - p) \log_2(1 - p) \,. \tag{1}$$

It is minimal and equal to zero for $p = 0$ and $p = 1$, while it is maximal and equal to one bit for $p = 1/2$ (uniform distribution). There is no information to be learned if an event is certain, and there is a maximum amount of information to be learned if probabilities are uniform. Low entropy means almost-certainty of the outcome before the experiment is performed, high entropy means high uncertainty about it, so that learning the outcome gives a lot of information.

e) The histograms show that the distribution is more spread for $p = 0.5$ than for $p = 0.1$, where the probability of obtaining 0 is larger than 0.6. Consistently, the entropy is larger for $p = 0.5$ than for $p = 0.1$.

2 Finite size effects on entropy

See Jupyter notebook.

3 Entropy as a measure of diversity in a microbial population

a) Why is the entropy of a population a measure of its diversity? If only one species is present, one frequency will be 1 and all others 0, and thus the entropy will be 0. If N species are present with identical frequencies $1/N$, the entropy will be

$$H(X) = -\sum_{x\in\Omega} P(x) \log_2[P(x)] = -N \times \frac{1}{N} \log_2\left(\frac{1}{N}\right) = \log_2(N) \,. \tag{2}$$

b) In Fig. 1, the diversity of the populations seems to first increase with time as more and more strains of different colors appear. This is due to mutations appearing. But afterwards, diversity seems to decrease, especially in panel (b), because some fitter strains with beneficial mutations increase in frequency due to natural selection. In panel (b), one of them seems to take over (the red one).

c) See Jupyter notebook.

d) See Jupyter notebook.

e) See Jupyter notebook. At the first time point, the entropy of the population is very small, almost zero, and accordingly, the histogram shows that one strain strongly dominates with a frequency close to 1. In fact this strain is the wild-type strain, at the first time point no mutation has appeared yet. At the last time point, the entropy is larger, and accordingly, the histogram shows some diversity, but nevertheless, a few strains dominate, as can be seen in Fig. 1(a) – purple strain, then orange and yellow ones etc.

f) See Jupyter notebook. The entropy of the population first steadily grows and then starts decreasing, except in the last epoch when it slightly increases again. It means that the diversity of the population first increases as mutants appear, and then it decreases as some fitter mutants take over. This corresponds to adaptation of the population to its medium. Note that there are some discontinuities in the entropy, as well as in the frequencies shown in Fig. 1: they correspond to the barcoding phases (shown in gray on Fig. 1) which perturb the population.

g) See Jupyter notebook. In this population too, the entropy of the population first steadily grows and then decreases. But here, it actually goes to almost zero at the end of the experiment, as one strain almost takes over (the red one in Fig. 1(b)), yielding a drastic decrease in diversity.