Problem set 6: Entropy

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Definition, interpretation and usual distributions

Recall that the entropy $H(X)$ of a random variable X is:

$$H(X) = -\sum_{x\in\Omega} P(x)\log_2[P(x)]\,, \tag{1}$$

where Ω is the set of possible values x that can be taken by the random variable X and $P(x)$ is a shorthand for $P(X = x)$.

a) Write your own function in Python that computes the entropy of a random variable X. What do you need to use about X?

b) Using the function you wrote, compute the entropy of a Bernoulli random variable with parameter $p = 9/10$, and then of a Bernoulli random variable with parameter $p = 1$. If you encounter a problem in the latter case, explain why. (If not, you can skip the rest of this question.) What should be the contribution to entropy of a state x such that $P(x) = 0$? Modify your function so that this case is handled well and check that you no longer encounter the previous problem.

c) Python has an entropy function in `scipy.stats`. Look it up and use it on the examples of the previous question. Check that you find the same results as with your own entropy function.

d) For 20 values of p equally spaced between 0 and 1, compute the entropy of a Bernoulli distribution with parameter p, and plot it versus p. Also do the analytical calculation. What is the maximum? For what value of p is it obtained? Comment on this.

e) Now consider the binomial distribution with parameters $N = 4$ and $p = 0.1$, and next, that with parameters $N = 4$ and $p = 0.5$. Using Python, plot their histograms and compute their entropies. Which of the two is larger? Discuss this by looking at the histograms. For this question you can use `binom` from `scipy.stats`.

2 Finite size effects on entropy

This problem was previously given as part of a numerical project for this class.

Consider a random variable X that can take K values, namely $0, \ldots, K-1$. The probability of the outcome $x = i$ is denoted by p_i. Assume that N samples of X are drawn, and denote by n_i the number of samples such that $x = i$ out of these N samples.

The number n_i can be seen as a random variable. Indeed, it depends on the particular set of N samples of X that are drawn. Imagine that you perform multiple times the procedure of drawing N samples and counting the number n_i such that $x = i$, then you will get various values of n_i.

a) What probability distribution does n_i follow? Justify your answer.

b) In all the following, we will assume that p_i is small and N is large but Np_i is finite. What probability distribution does n_i then follow? What are its mean and its variance? Consider the frequency $f_i = n_i/N$ of observing $x = i$. What are its mean and its variance (again over multiple draws of N samples of X)?

c) Let us define δ_i such that $f_i = p_i + \delta_i$. What is the mean of δ_i? What is its variance?

d) Let $\hat{H}(X)$ be the estimate of the entropy of X when using frequencies f_i instead of probabilities p_i, and let $\langle \hat{H}(X) \rangle$ be its average value over multiple draws of N samples of X. Using $f_i = p_i + \delta_i$ for each i, and assuming $|\delta_i| \ll p_i$, show that

$$\langle \hat{H}(X) \rangle \approx H(X) - \alpha \frac{K}{N}, \tag{2}$$

where $H(X)$ is the true entropy of X, computed using probabilities p_i, and α is a constant. Give the value of α.

If something goes wrong in your calculation or if you are not sure of your result, please admit Eq. 2 and use $\alpha = 0.7$ in what follows.

e) Comment on Eq. 2: what issue do we face when estimating entropy from a finite number N of observations (or samples)? [1]

f) Consider a random variable X that can take integer values from 0 to 9 with a uniform probability distribution. Using Python, for each value $N = 20, 30, 40, 50, 60, 75, 100$, construct 50 sets of N samples of this random variable. In each case, compute $\hat{H}(X)$, the entropy of X using the associated frequencies. You should thus have 50 estimates of $\hat{H}(X)$ for each value of N.

g) Using this data, compute the mean and the standard deviation of $\hat{H}(X)$, the estimated entropy of X, for each value of N. Produce a table containing the values of N in its first column, those of $1/N$ in the second one, those of $\langle \hat{H}(X) \rangle$ in the third one, and those of the standard deviation of $\hat{H}(X)$ in the fourth one.

h) Plot the mean estimated entropy $\langle \hat{H}(X) \rangle$ versus $1/N$, and show error bars with widths given by the standard deviations of $\hat{H}(X)$.

i) What is the exact value of the entropy of X? Plot it as a horizontal dashed line together with your previous plot.

j) Also plot the line given by Eq. 2 on the same graph. Comment.

k) Based on the plots you made, propose a method to estimate the actual value of the entropy of the distribution, if the exact value is unknown and you only have access to a given maximal number of draws of X, for instance 100 of them.

3 Entropy as a measure of diversity in a microbial population

In ecology and evolution, it is important to characterize the diversity of a population. For large populations, the frequencies of each species or strain can be used instead of probabilities to estimate the entropy of the population, which is a measure of diversity. Here, for simplicity, we will ignore the sampling problem studied above.

In Ref. [2], genetic barcodes were used to label subpopulations of a laboratory population of yeast. The frequency of each subpopulation was followed by tracking these barcodes and the mutations acquired by each subpopulation were also analyzed using sequencing. Due to selection and genetic drift, some subpopulations may grow and others disappear, and this dynamics can be followed using the barcodes. However, if e.g. one subpopulation takes over, barcodes are no longer useful to track the composition of the population and the dynamics of each lineage afterwards. To circumvent this issue, the population was re-barcoded at regular intervals. Barcoded lineages sharing the same beneficial mutations and the same ancestry were then assembled together and constitute a "clone", which we will here treat as a strain.

a) Why is the entropy of a population a measure of its diversity? You can analyze the case where only one species is present and the case where N species are present with identical frequencies.

b) Looking at Fig. 1, how do you think the diversity of the populations evolves with time? Why?

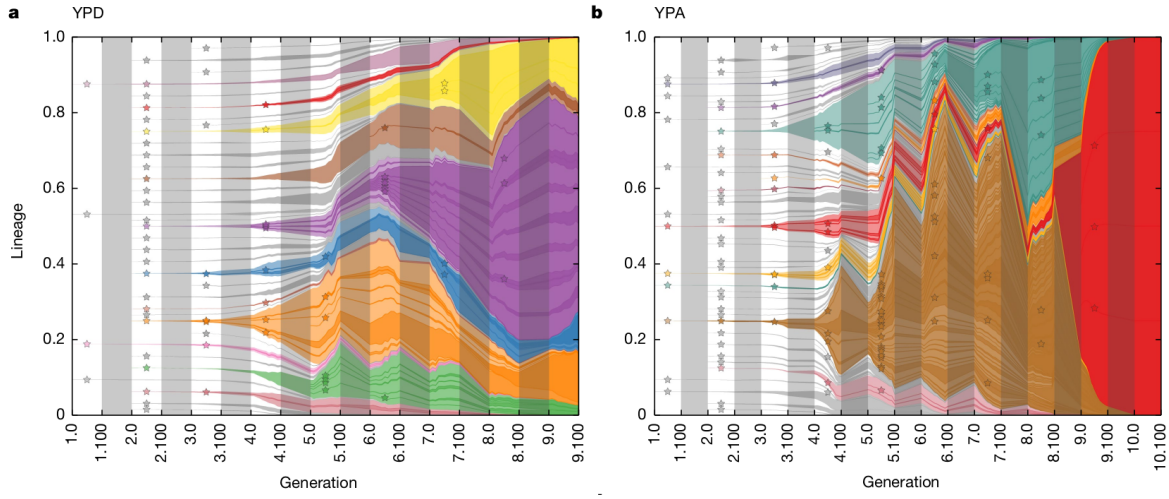

Figure 1: Stacked frequencies of subpopulations in a population of yeast versus time. Each color and nuance corresponds to a different clone or strain. New barcodes are added in the gray phases, and a new epoch starts at the end of each gray phase. Time is expressed in terms of epoch and generation (for example, 4.100 refers to generation 100 of epoch 4). Stars show the epochs where a beneficial mutation was acquired by a strain (and are shown in the same color as the corresponding strain). A large height (e.g. consider the purple strains at the right of panel (a)) means a large frequency of these strains. (a): Yeast in a rich medium (YPD). (b): Yeast in the same rich medium with added acetic acid (YPA). *Illustration from Ref. [2].*

c) In Python, load the data for Fig. 1, which is in the file `Data7.csv`. To visualize the structure of the table, you can first use `read_csv` from `pandas`, but then, for simplicity, it is easier to use `genfromtxt` from `numpy` to load the numbers in the file, e.g. via `my_data=np.genfromtxt('Data7.csv', delimiter='\t')`.

d) The frequency of each strain in the population in the first experiment (corresponding to Fig. 1(a)) is contained in `my_data[1:74,8:107]`. Each row is a different strain, each column is a different time point. Check that frequencies sum to one at all times (normalization).

e) Plot a histogram of the frequencies observed at the first time point and at the last one. Calculate the entropy of the population at the first time point and at the last one. What do you observe? What do you think happens at the first time point?

f) Calculate the entropy of the population at each time point, using a loop, and plot it versus the time index. How does the entropy of the population evolve in time? What does it mean about the diversity of the population? Why does this happen?

g) Repeat the 3 questions above for the population in the second experiment (corresponding to Fig. 1(b)), which is contained in `my_data[74:199,8:117]`. Compare the two populations.

References

[1] Bialek, W. *Biophysics: Searching for Principles*. Princeton University Press, 2012.

[2] A. N. Nguyen Ba, I. Cvijović, J. I. Rojas Echenique, K. R. Lawrence, A. Rego-Costa, X. Liu, S. F. Levy, and M. M. Desai. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. *Nature*, 575(7783):494–499, 11 2019.