Problem set 9: Maximum likelihood

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Estimating a diffusion coefficient

Consider a micrometer-size particle suspended in water. It undergoes Brownian motion as water molecules constantly hit it due to thermal fluctuations. Consider the displacement Δx, Δy of this particle in two dimensions between two successive observations under the microscope. The probability distribution function of these displacements is a Gaussian with zero mean:

$$p(\Delta x, \Delta y|V) = \frac{1}{2\pi V} \exp\left[-\frac{\Delta x^2 + \Delta y^2}{2V}\right] , \tag{1}$$

a) Assume that N measurements of displacements Δx, Δy of this particle are performed. They can be considered independent. What is the joint probability distribution function of the dataset $(\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2, \ldots, \Delta x_N, \Delta y_N)$ at a given V? Why can it be interpreted as a likelihood function? Express its logarithm. Why is it easier to work with it?

b) Maximize the likelihood obtained above, to obtain the maximum-likelihood estimate of V.

c) Using Python, load `Data10.npy`, which contains a list of measurements of $\Delta x_i, \Delta y_i$ expressed in micrometers, and coming from Jean Perrin's historic experiments about Brownian motion. Make a scatter plot of this data. Compute the maximum-likelihood estimate of V and of the standard deviation $\sigma = \sqrt{V}$ from this data.

d) The quantity $D = V/(2T)$, where T is the elapsed time between two successive observations under the microscope, is the particle's diffusion coefficient. Evaluate D, given that T was 30 seconds in Jean Perrin's experiment.

Note that the diffusion coefficient D is proportional to $k_{\mathrm{B}}T$, which reflects the fact that diffusion arises from thermal fluctuations.

2 Counting fluorescent molecules

Suppose that you look through a microscope at a sample containing some fluorescent molecules. They are individually visible, and they drift in and out of your field of view independently of one another. They are few enough in number that you can count how many are in your field of view at any instant. You make $K = 15$ measurements, obtaining the counts $n =$ 19, 19, 19, 19, 26, 22, 17, 23, 14, 25, 28, 27, 23, 18, and 26.

a) As a model, assume that the numbers above were drawn from a Poisson distribution

$$P(n) = e^{-\lambda}\frac{\lambda^n}{n!} , \tag{2}$$

where λ is the mean value of n. If this is true, how should the variance compare to the mean of the data? Using Python, check whether this is approximately true in the data above, and comment.

b) Write down the likelihood of the data $(n_1, n_2, \ldots, n_K)$ given the model with mean λ, then write down the log-likelihood. Find the maximum-likelihood estimate of λ and comment.

c) Using Python, plot the likelihood of the data given the model versus λ. Note that the likelihood may take huge numerical values, yielding an error. If this happens, subtract the maximum value taken by the log-likelihood (computed analytically from the previous question) from the log-likelihood, and take its exponential: like this, the new function will have a maximum value of one. Given your plot, what is the range of credible values of λ?

3 Luria-Delbrück experiment, revisited

Fig. 1 shows experimental data from a total of 87 trials of the Luria-Delbrück experiment. Specifically, the number m of phage-resistant bacteria that appeared in each replicate culture was counted, and here a histogram of m across 87 cultures is shown. The figure also shows an exact evaluation of the expected distribution assuming the Lamarckian hypothesis, as well as a simulated evaluation assuming the Darwinian hypothesis.

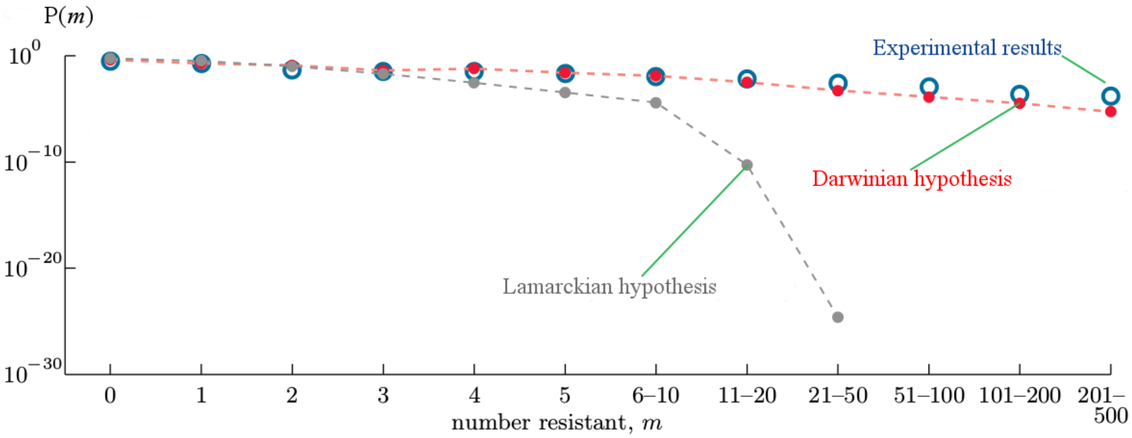

Figure 1: **Comparisons of the predictions from both hypotheses to data.** Empty blue markers: Data from Luria and Delbrück's article. Gray markers: fit to data under the Lamarckian hypothesis, using the Poisson distribution. Red markers: fit under the Darwinian hypothesis, using simulations of the jackpot distribution. A semilogarithmic scale is employed to better visualize the small probabilities associated to large m. Note that Luria and Delbrück combined the data for high mutant numbers m by lumping several values together, as indicated in the x-axis label. Here, when a bin contains k different values of m all lumped together, its count has been divided by k, so that the markers approximate the probabilities $P(m)$ for individual values of m. *Reproduced from Ref. [1] with modifications.*

a) Show that the logarithm of the ratio of likelihoods for the two models can be expressed as

$$\log\left[\frac{P(\text{data}|\text{model})}{P(\text{data}|\text{model}')}\right] = \sum_m n(m)\left[\log P(m|\text{model}) - \log P(m|\text{model}')\right] , \tag{3}$$

where $n(m)$ is the number of replicates of the experiment where m resistant bacteria were observed.

b) For each value of m, and using the logarithm with base 10, how can the value of the difference $\left[\log P(m|\text{model}) - \log P(m|\text{model}')\right]$ be read directly from the plot in Fig. 1? Estimate a lower bound for the logarithm of the ratio of likelihoods for the two models from the information in the plot. This does not need to be precise, and thus you should not need very precise estimates of the various quantities, but if you wish, you can go back to `Data3.npy` to get the values of $n(m)$.

c) Assume that you initially thought the Lamarckian hypothesis was five times more probable than the Darwinian. What would you conclude after the experiment?

4 Additional problem: Luria-Delbrück experiment again

This problem was previously given as part of the final exam of this class.

a) Under the "Lamarckian" hypothesis, mutations giving phage resistance occur in response to exposure of the bacteria to phage. Assume that they occur with a probability μ in each bacteria exposed to phage. What is the probability distribution of the random variable X characterizing whether one bacteria is resistant or not?

b) Now consider that a culture containing N bacteria is exposed to phage. For now, do not assume that μ is small. What is the probability $P(m)$ that m of them become resistant under the "Lamarckian" hypothesis? What is the name of this probability distribution?

c) What does this distribution become if $\mu \ll 1$ and $N \gg 1$ but $N\mu = \lambda$ is of order one? Give the name of this probability distribution and the corresponding expression of $P(m)$ as a function of m and λ. What are the mean and variance of this distribution? Make the calculation explicitly for the mean.

d) Explain what is different for the probability distribution of m under the "Darwinian" hypothesis where phage resistance mutations can be acquired randomly at any time during growth of the bacterial cultures (before exposure to phage).

e) In a recent study [2], the original data from the Luria-Delbrück experiment was reanalyzed. Recall that in this data, the number m of resistant bacteria is counted in many replicate cultures. The recent study addressed the extra possibility of a "Combined" model according to which both Darwinian and Lamarckian mechanisms coexist. It involves two parameters, the Darwinian mutation probability μ_D and the Lamarckian one μ_L. What model do we recover if $\mu_D = 0$? If $\mu_L = 0$? In this light, can the Combined model fit the data less well than the pure Darwinian or the pure Lamarckian models?

f) Write down the link between prior, posterior and likelihood for a given model and some given data. Clearly indicate what term is the prior, what term is the posterior and what term is the likelihood in the equation. What theorem did you use for this?

g) To compare two models, for a fixed dataset, what terms of the relationship above should be compared? Under what condition can this reduce to a maximum likelihood analysis?

h) What is the drawback of the Combined model compared to the pure Darwinian or the pure Lamarckian models? In what term of the relationship you just wrote could this be incorporated? We will not focus on this part anymore in this problem, but this aspect is important in practice.

i) The recent study reexamined in detail the results from two specific experiments described in the original Luria-Delbrück paper. These experiments, called experiment 22 and experiment 23, are those for which the number of analyzed cultures (or replicates) was the largest, among all experiments. Why is it important to have many replicates in a maximum likelihood analysis? Explain.

j) For experiment 22, the maximum likelihood fit obtained for the Combined model is shown in Fig. 2, with $\mu_L = 4 \times 10^{-10}$ and $\mu_D = 1.8 \times 10^{-9}$. Does it represent well the data? Comment.

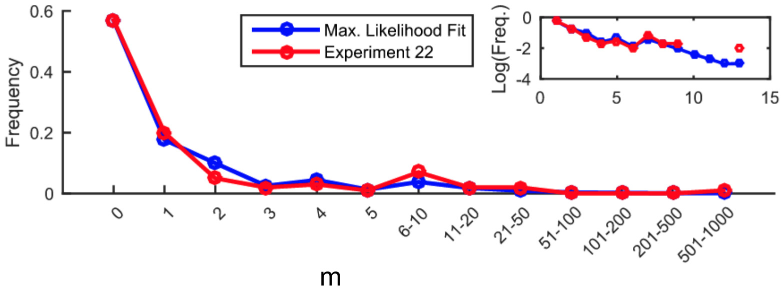

Figure 2: **Luria–Delbrück experimental data for experiment 22 and maximum likelihood fit of the Combined model.**

k) For experiment 23, the maximum likelihood fit obtained for the Combined model is shown in Fig. 3, with $\mu_L = 0$ and $\mu_D = 4.4 \times 10^{-9}$. Does it represent well the data? Comment.

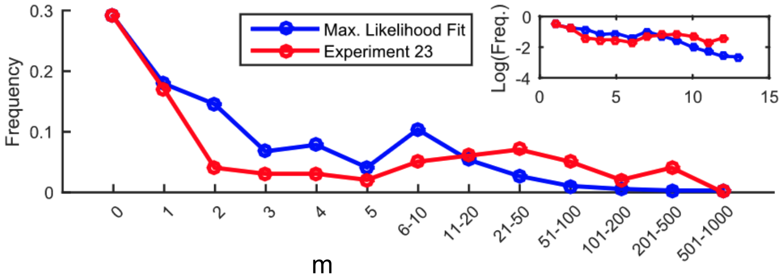

Figure 3: **Luria–Delbrück experimental data for experiment 23 and maximum likelihood fit of the Combined model.**

l) Based on these two results corresponding to experiments 22 and 23, do you think the Combined model can be excluded? What about the Lamarckian model? And the Darwinian one? What would you recommend to gain more insight on the comparison of the three models considered here?

References

[1] Nelson, P. *Physical models of living systems.* W. H. Freeman and company, 2015.

[2] C.M. Holmes, M. Ghafari, A. Abbas, V. Saravanan, and I. Nemenman. Luria-Delbrück, revisited: the classic experiment does not rule out Lamarckian evolution. *Phys. Biol.*, 14(5):055004, 2017.