

Lecture 1

Sayan Mukherjee

Preliminaries

The course is taught mainly in a case studies framework. The exceptions will be a few breaks to go over important methods or tools used in inference for a variety of statistical models in biology, namely expectation-maximization (EM) and Markov chain Monte Carlo (MCMC).

For a variety of biological problems we will proceed through the following steps:

1. Formalization of the biological problem – Statisticians call this modeling, the goal of this part is to come up with a *likelihood*
 - a. What are the relevant variables to model the biological question ?
 - b. Do we have the correct type of data to correspond to these variables ?
 - c. Which variables are (un)observable ?
 - d. What is our model of noise in the system ?
 - e. What is our model of uncertainty in the system ?
 - f. What are the parameters in our statistical model ?
 - g. Give our model what biological assumptions have we made ?
2. Estimating parameters from data – Statisticians call this inference (more specifically parameter estimation)
 - a. What is the uncertainty in my parameter estimates ?
 - b. Do these parameter estimates make any biological sense ?
 - c. How well does my data fit my modeling assumptions ?
 - d. If I fit many models how should I think about my results, which model should I pick ?
 - e. Does the model and parameters I chose fit future data or just explain what I have already seen ?
3. Answering the original biological questions using the inferred parameters and model(s).

An example

In 1894 Karl Pearson was part of a debate on whether inheritance of traits is continuous – the position of the Darwinian group – or particulate – the position of the Neo-Mendelian. The resolution of this debate resulted in what is called the “modern synthesis.”

In the paper “Contributions to the Mathematical Theory of Evolution” Pearson developed Gaussian mixture models to address this question. The trait examined is the breadth of the carapice (forehead) of crabs. We will reexamine this work.