HGEN 48600 Lecture Notes

Arjun Biddanda

January 22, 2016

Introduction

The primary theme of todays lecture is on likelihoods, likelihood ratios, and their interpretation. The writeup here *heavily* samples from the excellent vignettes found here

Example : DNA Barcoding of Poached Elephant Tusks

Elephant ivory is a commonly poached item, and we would like to determine which particular environment the elephants are coming from. The elephants cna either come from (1) the savannah or (2) the forest. We denote the following two models and data from an elephant tusk:

M_s : Tusk comes from a savannah elephant

M_f : Tusk comes from a forest elephant

Marker	Allele	f_S	f_F
1	1	0.4	0.8
2	0	0.12	0.2
3	1	0.21	0.11
4	0	0.12	0.17
5	0	0.02	0.23
6	1	0.32	0.25

f_S and f_F represent the allele frequency of the "'1" allele in the savannah and forest elephant populations respectively. The likelihood of the model

M_S can be defined as the probability of the data being generated under the model.

$$\begin{aligned}
L(M_S) &= P(Data|M_S) \\
&= \prod (f_S)^x \times (1 - f_S)^{1-x} \\
&= (0.4)(1 - 0.12)(0.21)(1 - 0.12)(1 - 0.02)(0.32) \\
&= 0.020399 \\
L(M_F) &= P(Data|M_F) \\
&= \prod (f_F)^x \times (1 - f_F)^{1-x} \\
&= (0.8)(1 - 0.2)(0.11)(1 - 0.17)(1 - 0.23)(0.25) \\
&= 0.0112 \\
LR(M_S/M_F) &= \frac{L(M_S)}{L(M_F)} \\
&= \frac{0.0204}{0.0112} \approx 1.8
\end{aligned}$$

We have defined the likelihood ratio as the ratio of the likelihood of the model M_S over the likelihood of M_F. We also note that the models M_S and M_F are *fully-specified*, that is to say that there are no free parameters in the model. We will explore unspecified models later on in the lecture.

Two distinct questions arise as a result of calculating this likelihood ratio : (1) how to interpret the likelihood ratio and (2) when can we claim that we believe a model over another model given a likelihood ratio?

Context of Individual Likelihoods

The purpose of a likelihood ratio is to examine the evidence (data) in the context of one model against another. As a brief toy example let us consider a fair coin tossed 100 times and landing with 50 heads and 50 tails. What is the likelihood of the model that this coin is a fair coin? $L(M_{fair}) = \left(\frac{1}{2}\right)^{100}$. However this very small number is simply a probability, and actually getting any set of 100 tosses with a fair coin results in this likelihood! Thus we can see that likelihoods are important by providing a context under which they can be interpreted, by comparing two models against each other.

Notational Things

- We often work in "log-space" since the individual likelihoods may be quite small. The Log-Likelihood Ratio (LLR) is defined as $log(LR)$
- In english we would say $P(Data|M_0)$ as "The likelihood under the model M_0"
- Sometimes semicolons are used to denote the data, and curly a "L" for the likelihood (i.e. $log(LR) = log\left(\frac{\mathcal{L}(M_S;D)}{\mathcal{L}(M_F;D)}\right)$)

Example : Continuous Measurement of Protein in Blood

Suppose that we a protein that is measured in the blood and we call this random variable X. We want to see if this protein varies in concentration according to disease or non-disease status We wish to test the following two models :

$$M_n : X \sim Gamma(0.5, 2)$$
$$M_d : X \sim Gamma(1, 2)$$

Where M_n is the model under a non-diseased state and M_d is a model under the diseased state. If we observe the data as $X = 4.02$ the likelihood ratio can be detemined as:

$$LR = \frac{f_{X|M_n}}{f_{x|M_d}}$$

. However we would only like to compare the density around the measurement we have actually obtained, so we will use a quick trick and assume the precision of the measurement of X to be ± 0.05 making $X \in [4.015, 4.025]$. Now that we have discretized this measurement we can integrate the respective conditional densities over this range, making the likelihood ratio :

$$LR = \frac{\int_{4.015}^{4.025} f_{X|M_n}}{\int_{4.015}^{4.025} f_{X|M_d}}$$

There are a couple of caveats to this approach as well that translate broadly to the calculation of likelihood ratios :

- The density function must be well-behaved in the integration bounds
- The data must be held the same between the models that we are comparing (transforming one and not the other is illegal!)
- If a likelihood ratio is 0 we can certainly say that

Different Support of Random Variables

Likelihood Ratios still work properly when we have different support for the models. Suppose we have a random variable X which corresponds to the roll of a standard 6-sided dice and the following two models:

$$M_6 : \text{All the dice rolls are a 6}$$
$$M_{fair} : \text{The dice is a fair dice}$$

We can imagine two scenarios from this : (1) when the roll is a 6 and (2) when the roll is not a 6. When X is a 6 we have a likelihood ratio of $LR = \frac{P(X|M_6)}{P(X|M_{fair})} = 1/(1/6) = 1/6$. However when $X \neq 6$ we can say that the likelihoo ratio is 0 since $P(X \neq 6|M_6) = 0$ and this is the numerator of our likelihood ratio.

Continuation of Disease Example

Let us assume that $Z_i = 1$ if patient is diseased, else $Z_i = 0$.

$$P(Z_i = 1|X_i = x) = \frac{P(X_i = x|Z_i = 1) \cdot P(Z_i = 1)}{P(X_i = 1)}$$
$$P(Z_i = 0|X_i = x) = \frac{P(X_i = x|Z_i = 0) \cdot P(Z_i = 0)}{P(X_i = 0)}$$
$$\frac{P(Z_i = 1|X_i = x)}{P(Z_i = 0|X_i = x)} = \frac{P(Z_i = 1) \cdot P(X_i = x|Z_i = 1)}{P(Z_i = 0) \cdot P(X_i = x|Z_i = 0)}$$
$$Odds_{Posterior} = Odds_{Prior} \times \text{Bayes Factor}$$

When the model is fully-specified the Bayes Factor is equal to the Likelihood Ratio. However the role of the prior odds also plays a large role in

our interpretation of the posterior odds. For instance if we believe that the disease is very rare, then we will have to have a much higher likelihood ratio in order to truly believe that we have the disease. It is very important to consider the prior odds of having the disease.

Likelihood Functions and Partially Specified Models

Let us review our elephant example now. But let us assume that we have sampled 100 elephants only from the savannah and look at their alleles at one particular marker. We obtain 40 samples that carry the "1" and 60 samples that carry the "0" allele. We then want to evaluate a particular model and its likelihood :

$$M_q : \text{Allele frequency of the 1 allele is q}, q \in [0,1]$$
$$\mathcal{L}(M_q) = P(Data|M_q) = q^{40}(1-q)^{60}$$

One way in which we can compare two potentially different values of q would be to look at their difference in log-likelihood units. We would then look at $log(\mathcal{L}(M_{q_1})) - log(\mathcal{L}(M_{q_2}))$. If we get a log-likelihood difference of 2 then we know that $LR = e^2 \approx 7.4$.