Bayesian Statistics II

Matthieu Vignes

April 2024

Comparing two models with a likelihood ratio

We introduce the concept of likelihood ratio (LR) for comparing two simple ("fully specified") statistical models - one of the most fundamental concepts in statistical inference. You "just" need to be familiar with basic concepts from probability, particularly the concept of a probability distribution.

Motivating Problem

[Technical Note: to simplify this problem, we assume that elephants are haploid. Which they are **not**.]

There are two subspecies of African Elephant: savannah and forest elephants, which differ in their genetic makeup. Interpol have seized an illegally-smuggled elephant tusk, and they want to know which subspecies of elephant the tusk came from. To try to answer this they collect DNA from the tusk and measure it at a number of locations ("markers" in genetics jargon) along the elephant genome. At each marker the DNA can be one of two types ("alleles" in genetics jargon), which for simplicity we will label 0 and 1. So the available data on the tusk might look something like this.

Marker	Allele
1	1
2	0
3	1
4	0
5	0
6	1

Interpol also have information on the frequency of each allele in each of the two subspecies - this was obtained by measuring the DNA of a large number of savanna elephants and a large number of forest elephants. We will use f_{S_j} and f_{F_j} to denote the frequency of `1` allele at marker `j` in savanna and forest elephants respectively (and since there are only two alleles, the frequency of the `0` allele is $1 - f_{S_j}$ and $1 - f_{F_j}$). Here is a table of this information.

Marker	fS	fF
1	0.40	0.80
2	0.12	0.20
3	0.21	0.11
4	0.12	0.17
5	0.02	0.23
6	0.32	0.25

The question we want to answer is: **Which subspecies of elephant did the tusk come from?** And a bonus question we will answer at the same time is: **How confident should we be in this conclusion?**

To get some intuition, let us examine the data at the first few markers. At marker `1` our tusk has the allele `1`. This allele is less common in savanna elephants than forest elephants (40% of savanna elephants carry this allele, vs 80% of forest elephants), so this observation seems to support the sample coming from forest. However, at the same time 40% of savanna samples carry this allele, so it remains plausible that the sample came from savanna.

Moving to marker `2`, our tusk has the allele `0`, which is more common in savanna (88%) than forest elephants (80%). And at marker `3` the tusk has allele `1` which is also more common in savanna (21% vs 11%). So, in contrast to marker `1`, the data at markers `2` and `3` are more consistent with the tusk coming from a savanna elephant than a forest elephant.

We could continue in this way, but the point should be clear: each marker contains some information. But not a huge amount. And sometimes the information points in different directions. By taking a statistical approach to this problem, we aim to quantify this kind of information, and to efficiently combine the information across markers to come to an overall conclusion.

(This is a very simplified version of a real problem: Interpol and other authorities want to know the origin of poached tusks to help focus efforts on curbing this illegal activity. In practice they are interested in much finer-level discrimination, and measure many genetic markers to get more information. See Wasser et al., 2007 and Wasser et al., 2008 for more details)

Solution

We can phrase this problem as a "model comparison" problem. We have data $X = x$ from our tusk, and we have two different models for how those data might have arisen: it could have been sampled from a savanna elephant, or it could have been sampled from a forest elephant. We will use M_S and M_F as shorthand for these two models. A key point is that these two models imply different probability distributions for X: some values of X are more common under M_S and others are more common under M_F.

Denoting the probability mass functions of these two distributions $p(.|M_S)$ and $p(.|M_F)$, and assuming the data at different markers are independent, these probability distributions are:

$$p(x|M_S) = \prod_j f_{S_j}^{x_j}(1 - f_{S_j})^{1-x_j}, \text{ and}$$

$$p(x|M_F) = \prod_j f_{F_j}^{x_j}(1 - f_{F_j})^{1-x_j}$$

where the values of f_S and f_F are given in the table above.

Note that a key feature of these models is that they are "fully specified". In other words, there are no unknown values in the probability distributions. Comparing fully-specified models is the simplest kind of model comparison, and so a good place to start in understanding the key concept of likelihood ratio introduced thereafter.

The likelihood ratio (LR)

The key idea to introduce here is that a useful summary of how strongly the data x supports one model vs another model is given by the "likelihood ratio".

The LR comparing two fully-specified models is simply the ratio of the probability of the data under each model. Notice that in saying the "probability of the data", we are assuming that the data and models involved are discrete. For continuous data and models the LR is the ratio of the probability densities of the two models evaluated at the data.

The name Likelihood ratio comes from the fact that the probability of the data under each model is called the "likelihood" for that model, when seen as a function of the parameters of the probability (distribution).

This is why it is recommended to say *likelihood for* the model, or *likelihood under* the model, rather than *likelihood of* the model, to help avoid confusing likelihood with probability.

That is, given data x the likelihood for a fully-specified (discrete) model M is defined as:

$$L(M) := P(x|M)$$

where $p(.|M)$ denotes the probability mass function for model M.

The likelihood ratio comparing two fully-specified (discrete) models M_1 vs M_0 is:

$$LR(M_1, M_0) := \frac{L(M_1)}{L(M_0)}.$$

Note that the likelihood for M depends on the data x. To make this dependence explicit the likelihood is sometimes written $L(M; x)$ instead of just $L(M)$.

Large values of $LR(M_1, M_0)$ indicate that the data are much more probable under model M_1 than under model M_0, and so indicate support for M_1. Conversely, small values of LR indicate support for model M_0, when compared to model M_1. How large is large? Wait, before that, a small numerical…

Example

Using the numbers in the tables above we can compute the likelihood and LR for M_S vs M_F:

```
x <- c(1, 0, 1, 0, 0, 1)
fS <- c(0.40, 0.12, 0.21, 0.12, 0.02, 0.32)
fF <- c(0.8, 0.2, 0.11, 0.17, 0.23, 0.25)
L <- function(f,x){prod(f^x*(1-f)^(1-x))}
LR <- L(fS, x) / L(fF, x)
print(LR)
```

```
## [1] 1.81359
```

So $LR(M_S, M_F; x)$ is `1.8135904`. This means that the data favor the tusk coming from a savanna elephant by a factor of about 1.8. This is a fairly modest factor – not large enough to draw a convincing conclusion. We will have more to say about interpreting LRs, and what values might be considered "convincing" later.

Note that we have deliberately focused on the likelihood ratio, and not the actual likelihood values themselves. This is because actual likelihood values are generally not useful - it is only the ratios that matter when comparing the models. One way of thinking about this is that the actual likelihood values are very context dependent, and so likelihoods from different data sets are not comparable with one another. However, the meaning of the likelihood ratio is in some sense consistent across contexts: $LR = 1.8$ means that the data favour the first model by a factor of 1.8 whatever the context.

The inverse likelihood ratio, and the log-likelihood ratio

Notice that, from the definition of the LR, the LR for M_0 vs M_1 is the just inverse of the LR for M_1 vs M_0. That is:

$$LR(M_0, M_1; x) = 1/LR(M_1, M_0; x).$$

So, for example if $LR(M_1, M_0) = 0.01$, then $LR(M_0, M_1) = 100$ and the data favours M_0 vs M_1 by a factor of 100.

For many reasons, it is common to work with the log likelihood ratio, $LLR := log(LR)$. Usually mathematicians work with logarithms base e, and we will use that convention unless otherwise stated. So, for example, if the $LR = 1.8$ then $LLR = log_e(1.8) = 0.5877867$.

Although the usual convention is to use log base e, it can sometimes be useful to work with logarithms base 10 to make the inverse logarithm operation easier for human calculation. For example, if I tell you that the LLR, base 10, is 3 then you can immediately tell me that the LR is 1000.

Exercises

1. You are playing a game with a friend. The friend has two six-sided dice, one blue and one green. The sides on the blue dice are numbered $1, 2, 3, 3, 3$ and 4. The sides on the green dice are labelled $1, 2, 2, 3, 4, 4$. He picks one of the dice without telling you, and rolls it 10 times, obtaining the results (data) $3, 3, 2, 3, 1, 2, 3, 3, 4, 3$. Looking at these results, does your intuition say anything in favour of the green dice or of the blue dice? Strongly or weakly? Phrase the problem as a model comparison problem. State your modelling assumptions, and compute a likelihood ratio. Does it support your intuition?

2. (Read the whole question before starting!)

    a. Perform the following simulation study. Simulate 1000 tusks (values of x) from each of the models M_S and M_F. For each simulated tusk compute the LR for M_S vs M_F, so you have computed 2000 LRs. Now consider using the LR to classify each tusk as being from a savanna or a forest elephant. Recall that large values for LR indicate support for M_S, so a natural classification rule is "classify as savanna if $LR > c$, otherwise classify as forest" for some threshold c. Plot the misclassification rate (= number of tusks wrongly classified/2000) for this rule, as c ranges from 0.01 to 100. What value of c minimizes the misclassification rate? [Hint: the plot will look best if you do things on the log scale, so you could let $log_{10}(c)$ vary from -2 to 2 using an equally spaced grid, and plot the misclassification rate on the y axis against $log_{10}(c)$ on the x axis.]

b. Repeat the above simulation study using 100 tusks from M_S and 1900 tusks from M_F. What value of c minimizes the misclassification rate? Comment.

```r
# n:  number of samples
# P:  a 2xR matrix of allele frequencies
r_simplemix <- function(n,P){
  R <- ncol(P) # number of loci
  z <- rep(0,n) # used to store population of origin of each individual
  x <- matrix(nrow = n, ncol = R) #used to store genotypes

  for(i in 1:n){
    z[i] <- sample(1:2, size = 1, prob = c(0.5,0.5))
    x[i,] <- r_haploid_genotypes(1, P[z[i], ])
  }
  return(list(x=x,z=z))
}
```

3. Consider now modifying our example above on the tusk to allow for errors in the data. Specifically, suppose that there is an error probability of 0.02 when measuring each marker: with probability 0.98 you observe the true x_j, but with probability 0.02 an error is made and you observe $1 - x_j$. Assume that errors occur independently at each marker `j`. Let M_S' and M_F' denote the models with this error process incorporated. Derive expressions for the likelihood for M_S' and M_F', and compute the LR for the example tusk data given here.

Summary

1. The likelihood for a model is the probability of the data under the model

2. Individual likelihood values are mostly irrelevant: it is likelihood ratios that matter

3. If the likelihood ratio for model 1 vs model 2 is r, then this means the data favours model 1 by a factor of r. (Or, if $r < 1$ then it means that the data favours model 2 by a factor of $1/r$)