# Comparing two models with a likelihood ratio

**Matthew Stephens**    *University of Chicago*

## Summary

This document introduces the concept of likelihood ratio for comparing two simple ("fully specified") statistical models, one of the most fundamental concepts in statistical inference.

## Pre-requisites

Be familiar with basic concepts from probability, particularly the concept of a probability distribution.

## Motivating Problem

**Technical Note:** To simplify this problem I have assumed that elephants are haploid, which they are not. If you do not know what this means you should simply ignore this comment.

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 $f_{Sj}$ and $f_{Fj}$ 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 $1 - f_{Sj}$ and $1 - f_{Fj}$). Here is a table of this information.

| marker | $f_S$ | $f_F$ |
|--------|-------|-------|
| 1 | 0.40 | 0.80 |
| 2 | 0.12 | 0.20 |

| marker | $f_S$ | $f_F$ |
|--------|-------|-------|
| 3 | 0.21 | 0.11 |
| 4 | 0.12 | 0.17 |
| 5 | 0.02 | 0.23 |
| 6 | 0.32 | 0.25 |

The question before us is: Which subspecies of elephant did the tusk come from, and 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 these two markers 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 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.etal.07 and @wasser.etal.08 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 $p(\cdot|M_S)$ and $p(\cdot|M_F)$, and assuming the data at different markers are independent, these probability distributions are:

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

and

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

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". This means 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 here.

**The likelihood ratio**

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

The LR comparing two fully-specified models is simply the ratio of the probability of the data under each model. (In saying the "probability of the data" here 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, as discussed in detail here)

The name Likelihood ratio comes from the fact that the probability of the data under each model is called the "likelihood" for that model. I recommended saying "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 $p(\cdot|M)$ denotes the probability mass function for model $M$. And the likelihood ratio comparing two fully-specified (discrete) models $M_1$ vs $M_0$ is defined as

$$LR(M_1, M_0) := L(M_1)/L(M_0).$$

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 $M_0$.

**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] 1.814
```

So $LR(M_S, M_F; x)$ is 1.8136. 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 favour $M_0$ 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 base 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)=0.5878.

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 3,3,2,3,1,2,3,3,4,3. Looking at these results, does intuition say that they support the green dice or 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 log10(c) vary from -2 to 2 using an equally spaced grid, and plot the misclassification rate on the $y$ axis against log10(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.

Note: it is good practice when writing computing code for a simulation study like this to separate the code into small functions each of which accomplishes a clearly-defined task. So, for example, you should write a function to simulate data, and another to compute the LR (possibly using the function L defined above!), and another to compute the misclassification rate, etc. It is also good practice to comment your functions: eg say what the input parameters are and what the output is. I strongly suggest using roxygen2 format. See Karl Broman's tutorial for example.

*DO NOT REPEAT THE SIMULATION STUDY BY COPYING AND PASTING CODE AND MODI-FYING A FEW NUMBERS! GET INTO THE HABIT OF DOING IT PROPERLY*

To be more explicit, at the top level your code should probably include a function something like this:

```
#' Simulation study
#'
#' Simulates data from forest and savanna tusks from prespecified frequencies, and uses LR to cl
#'
#' @param nS number of savanna tusks to use
#' @param nF number of forest tusks to use
#' @param fS allele frequencies of savanna elephants
#' @param fF allele frequencies of forest elephants
#' @param lc vector of log(c) values of LR cutoff to use
#' @param seed set a random seed using set.seed() for reproducibility
#'
#' @return Plot of misclassification rate against threshold
#'
#' @example simulation_study(1000, 1000, c(0.40,0.12,0.21,0.12,0.02,0.32), c(0.8,0.2,0.11,0.17,0
#'
#'
simulation_study(nS, nF, fS, fF, lc, seed){
...
}
```

Then you can just call this simulation study function twice to complete the assignment.

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

Repeat after me:

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 x, then this means the data favour model 1 by a factor of x". (Or, if x <1 then it means "the data favour model 2 by a factor of 1/x"")

## References