Frog calls 4 : modelling

On the first day, we used dice to simulate data for detection of frogs at 10 ponds known to be inhabited by frogs. The observations can be described (or "modelled") using the binomial distribution. We later looked at the concept of ' likelihood' and the maximum likelihood estimate (MLE) for the probability of detection, $\hat{P}$ (detect).

The calculations for this session are in the Excel file "Frogs4_models.xls"; the text here refers to Excel 2007 commands, but other software has similar functionality.

Note down the number of ponds where you heard frogs calling during the first simulation (we'll call this x_1) : x_1 =

Simulating data for a second visit

We'll again use a polyhedral die with spots on some of the sides. If the die lands with a spot on the upper-most side, the frogs were calling; if there's no spot, they were silent.

Roll the die 10 times and record the results (☑ = calling, ☒ = silent):

How many times did you hear frogs calling? (We'll call this x_2.)

x_2 =

When I did this, I got x_1 = 6 and x_2 = 3. You will probably get different values.

You may get the same values for x_1 and x_2; that's not much fun, though, so repeat the second simulation to get different values.

With different values for x_1 and x_2, we can consider two possible models:

Model 1 : the true detection probability was the same on both occasions, the difference is due to sampling error.

Model 2 : the detection probability was different on the two occasions.

Model 1

Model 1 has a single value for *p*(detect), and we can use the simple maximum likelihood estimator, $\hat{P}$ *(detect)* = $(x_1 + x_2) / 20$. With my data, that gives $\hat{P}$ (detect) = (6 + 3) / 20 = 0.45

Open the Excel file "Frogs3_models.xls".

We'll see how this works with my results first; later you can insert your results into the spreadsheet and see your results.

Insert my results for x_1 and x_2 in cells C1 (6) and E1 (3). The number of trials is 10; put that in cell C2.

You will see the same numbers (6 and 3) in line 5 (and in line 12 too). The total number of detections (9) appears in cell D5, just under the heading "overall". We use the overall number of detections to calculate a single estimate for *p*, which we'll put in cell E3: =D5/(2*C2). Look at the formula and make sure you understand it.

Hint: use the Trace Precedents function (on the Data tab in Excel, Tools > Detective in Calc) to see where the results come from.

We don't need to calculate likelihoods for all possible values of *p*, but we do need the maximum value of the likelihood, ie, the value when $p = x/n$. Remember that likelihood is just the probability of getting the data we got with the given value of *p*.

In cell B6, calculate the probability of getting the data (ie, 6 detections) out of 10 with our estimate of p. Use BINOMDIST. Use cell references, not numbers, so that everything works when we change x1 and x2 in row 1 later.

Do the same for the second occasion in cell C6.

The overall likelihood is just the product of the individual likelihoods.

In cell D6, calculate the product of B6 and C6.

I get the overall likelihood to be 0.0265; you should get the same if you are using my data.

Now we'll turn to Model 2, and see how the overall likelihood compares.

Model 2

This model has two different values of *p*(detect), one for the first occasion and another for the second occasion (p_2). In the spreadsheet, you will see the data appear in line 12, and estimates for p_1 and p_2 are calculated in line 10 using the maximum likelihood estimator, x/n. Make sure you understand where these come from.

Use BINOMDIST again to calculate the likelihoods in cells B13 and C13; remember to use different probabilities for each. Then calculate the overall likelihood in cell D13.

I got an overall likelihood of 0.0669.

Comparing the models

Comparing the two models, we see that the likelihood for model 2 is higher than for model 1, which means that it is a closer fit to the data we collected. If we use the "maximum likelihood" idea, we would say that Model 2 is better than Model 1.

Adding more parameters to the model will *always* increase the likelihood. It always improves the fit to the data you already have, but it does not mean predictions based on the model are better. If you want to predict the probability of detecting frogs in the future, a model with fewer parameters will likely be better.

Hirotugu Akaike had a similar problem in his research on cement, and developed a criterion for the best model, based on the log of the likelihood and the number of parameters estimated from the data:

Akaike's Information Criterion (AIC) = 2 x no. of parameters – 2 x log(likelihood)

The model with the lowest AIC is the best model.

Calculating AIC

In cell D7 calculate the log of the overall likelihood for Model 1. In Excel and Calc use LN() to get the natural logarithm.

The number of parameters estimated from the data is 1 for Model 1. Calculate the AIC for Model 1 in cell D9.

Do the same to calculate the AIC for Model 2, which has 2 parameters estimated from the data.

Calculate the difference in AICs in cell D18.

With my data, I get AICs of 9.26 (Model 1) and 9.41 (Model 2). Model 1 has the lowest AIC, so it is the better model. Still, the difference in AICs is quite small, only 0.15.

Analysing your own data

Put your own values for x1 and x2 into the top row of the spreadsheet. If all goes well, all the values will change and you will get the AICs for each model based on your data. Which model is best for your data?

Compare your results with other participants: How many found Model 2 to be better than Model 1? What sort of results (x1 and x2) led to that conclusion?

The true values (which we know in this case) are very different: $p_1 = 0.75$ and $p_2 = 0.33$, so Model 2 is the true model. But with small samples like this ($n = 10$), there is still uncertainty about the conclusion. The size of the difference in AIC gives a clue to how certain we are about the result, and we'll look at that next. In fact we can go further with AIC and calculate probabilities for each of the models. We will see how to do that later.

Main points

- A **model** is a simplified description of a real-world phenomenon, often expressed as a mathematical relationship.
- If we use Maximum Likelihood Estimation to estimate the **model parameters**, we can use the value of the likelihood to compare models.
- Because the likelihood is usually extremely small, we use the logarithm, the **log likelihood**.
- The model with the highest likelihood is the best fit to the *data*, but a model with fewer parameters may be a better fit to the underlying *phenomenon*.
- **AIC** balances number of parameters and log likelihood; the best model has the lowest AIC.