Lecture # 5 - Monday, Aug 30th

In this lecture I reviewed the previous lecture 4, and then reviewed the final point of lecture #3: Hennig's logical approach to tree inference is valid as logic, but we have a hard time using it because our data does not fulfill the requirements of logical premises. Essentially, we can never know that our homology statements about character states are true in the strictest sense. If two taxa that show state 1 for some character, we cannot *know* that the common ancestor of must have had that state – we must admit that there is some chance of convergence that we did not detect when coding the data.

When we introduce uncertainty we have to move to statistical inference (or perhaps to "fuzzy logic", but we are not going to discuss that in this course).

I gave an example using the method of moment estimator. The purpose of example is to outline the basic structure of a statistical argument: We have set of models (or hypotheses) for how the world behaves. We can imagine generating data from those models – so the models make probabilistic statements about what we would *expect* the data to look like.

We can conduct inference by seeing which model predicts an outcome that is most similar to the data that we observe.

Later we'll talk about statistical testing in which we ask the question: is it plausible that this model could have generated this data set?

Lots of approaches fall under the umbrella of valid statistical inference procedure.

Example

For example, consider a case in which someone has 5 dice. Each one could be either fair die or it could have 1 on each side. Imagine that he roll all five and reports the results as: 1, 3, 1, 2, 3.

We would like to infer the number of dice that are actually one-sided (throughout these notes I'm going to use "one-sided" as shorthand for a die in which all six sides have a one on them).

It is cumbersome to deal with the entire dataset, $X = [1, 3, 1, 2, 3]$. So we could represent the dataset with a simple summary statistic. For example we could use the sample mean $\overline{X}$. So our data can be summarized as $\overline{X} = 2$.

There are 6 distinct models that we could use: M_0 states that there are 0 one-sided dice and 5 fair dice; M_1 states that there is 1 one-sided die and 4 fair dice; ... M_5 states that all 5 die are one-sided. We could view these as all the same model, and treat the number of one-sided dice as a free, discrete parameter in that model that is the subject of inference. For this problem it does not matter whether we view this as an example of finding the best-fit model or the best parameter value.

Without looking at the data we can evaluate what type of data that the models could generate. For example, we could consider the experiment of drawing a set of rolls from each these models. We can say what the expected value of the mean would be.

Formally the expected value of some function, f, of a random variable, X, over a probability distribution, p, which describes the probability of each possible value of x from the set of all

possible values (this set of all possible values is denoted $\mathcal{X}$) can be expressed as:

$$\mathbb{E}_p(f(X)) = \sum_{x \in \mathcal{X}} f(x)p(x) \tag{1}$$

The function that we are interested in is the arithmetic mean $f(X) = \overline{X}$. The role of the model is to specify what types of data sets are common – to assign a probability to every possible outcome.

There are lots of possible outcomes of rolling 5 dice. In fact there are $6^5 = 7776$ possible datasets.

Fortunately there are some tricks that we can use. For a single fair die the expectation of the value is 3.5:

$$\begin{aligned}
\mathbb{E}_{\text{FAIRDIE}}(X) &= \sum_{x \in \{1,2,3,4,5,6\}} x P_{\text{FAIRDIE}}(x) && (2)\\
&= \sum_{x \in \{1,2,3,4,5,6\}} x \left(\frac{1}{6}\right) && (3)\\
&= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) && (4)\\
&= \frac{21}{6} && (5)\\
&= 3.5 && (6)
\end{aligned}$$

The expected value for a one-sided die is 1 (unsurprisingly). This is obvious, but do note that the long way of doing it always works:

$$\begin{aligned}
\mathbb{E}_{\text{ONESIDED}}(X) &= \sum_{x \in \{1,2,3,4,5,6\}} x P_{\text{ONESIDED}}(x) && (7)\\
&= 1\,(1) + 2\,(0) + 3\,(0) + 4\,(0) + 5\,(0) + 6\,(0) && (8)\\
&= 1 && (9)
\end{aligned}$$

Because we know that the mean of five rolls just a sum of the contributions of each die divided by five (the total number of rolls), we can just sum the expectations for each roll and then divide by 5. For example in the M_0 model we would add 3.5 five times and then divide by 5, while for the M_2 model we would add 1 twice and 3.5 three times before dividing by 5. These considerations, and general observations of the highest and lowest possible means for any trial give us the table of predictions for each model shown in Table (2).

In principle we can do statistical inference whenever models make different predictions about the outcome. Looking at the predictions, we could say that a way to test the models would be to collect lots of data on the mean value from a set of 5 rolls. If you did lots of experiments, you could look at the largest mean that you ever observed. Because the models differ in terms of their predictions about the largest mean, you could use this experiment as a basis of preferring the model that best matches the data.

Based just on the calculations shown in Table (2), we would *not* collect lots of data and just keep the minimum value of the mean. Doing a huge number of trials and just recording the minimum is very likely to give you the value 1 – which is predicted by all of the models, so you have no way of discriminating between them.

Table 1: Some predictions of the 6 models

Model	$\mathbb{E}(\overline{X})$	$\max \overline{X}$	$\min \overline{X}$	The smallest number of 1's observed in any set of rolls
M_0	3.5	6	1	0
M_1	3.0	5	1	1
M_2	2.5	4	1	2
M_3	2.0	3	1	3
M_4	1.5	2	1	4
M_5	1.0	1	1	5

We can base an estimation procedure on performing lots of experiments and getting a grand mean – the models differ in their predictions about the expected value of the mean so we can use the mean to discriminate.

The models also differ in their prediction about the fewest number of die that display 1 in any set of rolls (if there are 2 one-sided dice, then you should always see at least 2 dice with 1).

So three potential bases of estimation appear to be: the "mean of the means", the maximum $\overline{X}$, and the fewest numbers of 1's in any trial.

Comparing estimators

We have just one trial. It had $\overline{X} = 2$ and two of the dice displayed a 1. Thus, We could view 2 as either the largest mean we have observed in our trials, or the mean over all trials, or the smallest number of 1's in any trial.

Notice that:

- if we base our estimation on the $\max \overline{X}$, then model M_4 comes closest to our observed value;
- if we base our inference on the mean of the means (the mean of $\overline{X}$ for each trial) then M_3 is the preferred model; and
- if we judged the outcome based on counting the number of die that display "1" in each set of rolls, then M_2 would be the preferred model.

Which form of estimation is "valid"? All three are, in a sense. If we gave these three estimation schemes enough data they would all get the right answer. In logical inference, you would never follow two valid sets of rules and arrive at conflicting inferences. But in statistical inference this can happen. We have to admit that there can be sampling error and we do not expect a statistical estimator to always return the truth.

Just because multiple estimators are valid in the sense that they would get the answer right if you could get rid of sampling error, does not mean that they are equally good. We do not define statistical estimators as "valid" or "invalid." There are a whole slew of properties of estimators that we use to evaluate them: is the estimator biased? how quickly does the bias disappear as we add more data? how precise are the estimates?

The behavior of estimators depends a lot on how efficiently they use information in the data and

how reliable that information it is.

If you think about it a bit then, I think that you'll realize that the mean of means estimator has an advantage over the other two in that it accumulates some information with each trial. Every time we roll the set of five dice we get a better estimate of the mean of means.

The other two estimators have an advantage in the sense that they can exploit some more informative aspects of the data. M_5 says that $\overline{X} > 1$ is *impossible*, and it is *impossible* to observe a roll with fewer than 5 dice showing 1. These are very strong statements. If we base our inference on either of these estimation schemes, then we would immediately know that M_5 is not only not the preferred model it can be rejected. From the information that I've given you, you would not be able to reject M_5 based just on the mean being different from $\mathbb{E}_{M_5}(\overline{X})$.[1]

Also notice that two of the estimators (those based on the maximum $\overline{X}$) seems to obscure the quantity that we are interested in (the number of 1's in an outcome) into the mean. Surely, if we want to estimate how many dice are one-sided, then knowing that a set of rolls came out as $[1, 1, 1, 2, 5]$ is more informative than knowing that mean was 2. If we just know that the mean was two we could have as few as 0 dice showing 1 (the outcome $[2, 2, 2, 2, 2]$) or as many as 4 dice showing 1 (the outcome $[1, 1, 1, 1, 6]$). If you know that a die is not showing 1, do you even care what value it shows? Perhaps in a statistical context (just like on facebook) you can have TMI (?).

So it appears that none of the three estimators discussed have potential positives and negatives when it comes to using the data to its fullest extent. It is not clear (at least not to me) which of these represents the best tradeoff[2]

Likelihood

There are lots of estimators to choose from. Fortunately, statistical theory makes some general recommendations so that we don't have to evaluate every form of estimator. For a large class of problems in which we have a model that we can express as a probability statement over different outcomes, then we should go an estimator that is based on the likelihood. Specifically, the "law of likelihood" states that all of evidence in favor of one parameter value (or "one model" or "one hypothesis") over another value is contained in the likelihood ratio (see http://en.wikipedia.org/wiki/Likelihood_principle).

Maximum Likelihood Estimators

For a great many problems, the maximum likelihood estimator is often the most efficient estimator (among the class of estimators that are not asymptotically biased[3]) and the maximum likelihood estimator (MLE) is consistent[4].

[1] In truth, one could also examine variance from repeated trials that one would expect in $\overline{X}$, and this would reveal the under M_5 we expect no variance – so there is more information in the moments than we are using here.

[2] My intuition would suggest that that, among these three forms of estimation, the best-to-worst ordering would be: "count the number of ones", then "use the mean of means", and then "use the maximum value".

[3] "Biased" estimators will tend to make an error in one direction – *eg.* For a numerical variable this might mean that they either overestimate or underestimate the true value. MLE's are often biased estimators, but the bias usually disappears as the sample size increases. Sometimes the bias can disappear slowly.

[4] Roughly speaking, "consistency" means that an estimator will converge to the right answer as the number of data points increases

Likelihood - definition

If X is the data, then the likelihood of a model M is:

$$\Pr(X|M)$$

In words it is "the probability of observing a data set exactly like the one that we observed if the data were being generated by the model M"

The likelihood is an assessment of how well the data fits the world that we have observed. And it is not just any assessment. According to the law of likelihood *all* of the evidence that the data has to bring to bear in deciding between the models is in the likelihood ratio.

How do we calculate a likelihood in this context? It involves a lot of basic probability calculations. Note that we are interested in the probability of observing a trial identical to the outcome that we actually observed. So we are going to make a probability statement about seeing $[1, 3, 1, 2, 3]$.

Let's take M_0 first. When we say that a dice is fair then we are saying that all six sides should come up with equal probability $(1/6)$, and that one roll is independent from every other roll. The events A and B are *statistically independent* from each other then:

$$\Pr(A, B) = \Pr(A)\Pr(B)$$

Thus,

$$\begin{aligned}
\Pr([1, 3, 1, 2, 3]|M_0) &= \Pr(1)\Pr(3)\Pr(1)\Pr(2)\Pr(3) && (10)\\
&= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) && (11)\\
&= \frac{1}{7776} && (12)
\end{aligned}$$

The calculations under M_5 are also easy because $\Pr(1) = 1$ for every roll, so $\Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = 0$

$$\begin{aligned}
\Pr([1, 3, 1, 2, 3]|M_5) &= \Pr(1)\Pr(3)\Pr(1)\Pr(2)\Pr(3) && (13)\\
&= (1)\,(0)\,(1)\,(0)\,(0) && (14)\\
&= 0 && (15)
\end{aligned}$$

Notice that the likelihood has "detected" that M_5 is impossible, and the model receives the lowest possible score.

Calculating the likelihood of the other model is a bit more complex because we have to consider all the combinations of which dice were fair and which were one-sided. I'll work through M_1 to demonstrate another rule of probability. Namely, if events A and B are mutually exclusive, then

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B).$$

The more general formulation is that

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A, B)$$

where we subtract off the probability that both A and B occur. Different sides of the die are mutually exclusive (more than one cannot occur in any single trial), thus $\Pr(A, B) = 0$ in this context so we don't have to worry about subtracting out that bit.

Before we look at the data we know that for a trial under M_1, the one-sided die could (with probability 1/5) be the first die listed, *or* it could (with probability 1/5) be the second, *or* the third ... Thus we are going to add several possible scenarios. If we use D_F to denote fair die and D_1 to denote a one sided die then:

$$\begin{aligned}
\Pr([1,3,1,2,3]|M_1) &= \left(\frac{1}{5}\right)\Pr(1|D_1)\Pr(3|D_F)\Pr(1|D_F)\Pr(2|D_F)\Pr(3|D_F)\ldots && (16)\\
&+ \left(\frac{1}{5}\right)\Pr(1|D_F)\Pr(3|D_1)\Pr(1|D_F)\Pr(2|D_F)\Pr(3|D_F)\ldots && (17)\\
&+ \left(\frac{1}{5}\right)\Pr(1|D_F)\Pr(3|D_F)\Pr(1|D_1)\Pr(2|D_F)\Pr(3|D_F)\ldots && (18)\\
&+ \left(\frac{1}{5}\right)\Pr(1|D_F)\Pr(3|D_F)\Pr(1|D_F)\Pr(2|D_1)\Pr(3|D_F)\ldots && (19)\\
&+ \left(\frac{1}{5}\right)\Pr(1|D_F)\Pr(3|D_F)\Pr(1|D_F)\Pr(2|D_F)\Pr(3|D_1) && (20)
\end{aligned}$$

This works out to:

$$\begin{aligned}
\Pr([1,3,1,2,3]|M_1) &= \left(\frac{1}{5}\right)(1)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\ldots && (21)\\
&+ \left(\frac{1}{5}\right)\left(\frac{1}{6}\right)(0)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\ldots && (22)\\
&+ \left(\frac{1}{5}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)(1)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\ldots && (23)\\
&+ \left(\frac{1}{5}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)(0)\left(\frac{1}{6}\right)\ldots && (24)\\
&+ \left(\frac{1}{5}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)(0) && (25)\\
&= \frac{1}{5(6^4)} + 0 + \frac{1}{5(6^4)} + 0 + 0 && (26)\\
&= \frac{2}{6480} && (27)\\
&= \frac{1}{3240} && (28)
\end{aligned}$$

M_2 is even more complex because we have to consider all of the arrangements of the 2 one-sided dice in the set of 5 rolls. There are $\binom{5}{2}$ of these. Recall that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. In this case it works out to 10 arrangements. Each would be expected to occur with a probability of 1/10 Also note that only one arrangement (with the one-sided dice as the first and third dice reported in the trial) is consistent with the data. The other nine arrangements will have factors of 0 in them, and so can

be ignored.

$$\Pr([1,3,1,2,3]|M_2) = \left(\frac{1}{10}\right)(1)\left(\frac{1}{6}\right)(1)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\ldots \quad (29)$$

$$= \frac{1}{10(6^3)} \quad (30)$$

$$= \frac{1}{2160} \quad (31)$$

Models M_3, M_4, and M_5 will each have a likelihood of 0 because each of them will trying to explain a non-zero roll as coming from a one-sided die. In summary, likelihood estimation gives the following likelihoods:

Table 2: Likelihoods of the six models

Model	$\Pr(X = [1,3,1,2,3]\|\text{Model})$
M_0	1/7776
M_1	1/3240
M_2	1/2160
M_3	0
M_4	0
M_5	0

Note that the MLE is M_2, but the the preference between models (as expressed by the likelihood ratio is not strong). A common (very rough) rule-of-thumb when looking at models that have the same number of parameters is to look at the MLE and all parameters (or models) within 2 log-likelihood units of it. $e^2 \approx 7.4$ so we'd want to consider all models that with likelihoods within a factor if 7.5 of the MLE. Note that M_0 and M_1 would both be in this set, so our data has not rejected either of them.

But our data does reject M_3, M_4, and M_5. The likelihood detects that these models are fundamentally inconsistent with the data.

Also note that, we do not have to tell the the estimation procedure that some of the information is essentially irrelevant. For all of the models $\Pr([1,3,1,2,3])$ would be the same as $\Pr([1,3,1,2,4])$ or any other arrangement with two 1's.

In summary, there is an estimation procedure (maximum likelihood) that is efficient because it makes use of all of the information in the data. It automatically "does the right thing" with irrelevant information in the data, so we don't have to worry about preprocessing our data so that irrelevant aspects are ignored. For most models, likelihood-based estimation results in consistent estimation of the true value of the parameter, and will be the most powerful estimator. While MLE's can be biased on small datasets, the consistency condition guarantees that the bias will disappear if we have enough data (but still subject to the *caveat* that for poorly-behaved models the MLE is not guaranteed to be a consistent estimator).