Homework #3

(due Monday, Feb 25th)

(1) A snail species has two different morphs, left-coiled and right-coiled. A researcher does a large experiment measuring the escape behavior of snails when confronted with a predator. She classifies each as 'run' or 'hide' (pull into shell). 50% of the snail population exhibit the run behavior. Among runners, 25% are left-coiled. Among hiders, 60% are left-coiled. Calculate the probability that a snail runs given that it is right-coiled.

(2) Let's revisit the analysis of data on the number of substitutions on two branches of a genealogy, where the two branches have the same time to the most recent common ancestor. We'll denote the data on counts of the number of changes is y_1 and y_2. Previously JKK worked through the case of using a Poisson distribution with an expectation of ut as a model for these data. He only used the number of sites with differences.

Now imagine that we know the number of sites that we have sequence from. Call that total M. Since mutations are rare M is typically much larger than $y_1 + y_2$. If we imagine that mutations are rare independent events that happen with the same rate at every site, then we can view the data on y_1 and y_2 as given M sites as samples from the binomial distribution. Let p be the probability of a change at one site on a single branch.

(a) What is the likelihood equation for p given y_1, y_2, and M trials?

(c) What is the formula for the MLE, $\hat{p}$?

(c) What is the MLE of p if $y_1 = 2$, $y_2 = 3$, and $M = 2000$?

(d) What is the log-likelihood at that point?

(e) Can you reject a null hypothesis that $p = 0.001$? (show the LRT test statisic and *df*)

(3) If you knew exactly where in the sequence of 2000 sites the 5 changes occurred, then you could calculate the probability without the binomial coefficient instead of the one you used in #2b. If you used a likelihood equation without the binomial coefficient, you would get a (Circle one answer for each part):

(a) Lower / Equal / Higher likelihood
(b) Lower / Equal / Higher log-likelihood
(c) Lower / Equal / Higher estimate of $\hat{p}$
(d) Lower / Equal / Higher LRT for the H_0 that $p = 0.001$

(compared to the values of these entities when you use a likelihood equation with the binomial coefficient).

Notes which cover the answers to question #2

connecting mutation rate to a probability of mutation

For closely related sequences, we can ignore "multiple hits" (two mutations to the same site that result in the number of differences being less than the number of mutations).

If we have a mutation rate we can use the Poisson where $\lambda = ut$.

$$\begin{aligned}\mathbb{P}(\text{a difference} \mid u, t) \approx \mathbb{P}(> 0 \text{ mutations} \mid u, t) = p &= 1 - \mathbb{P}(0 \text{ mutations} \mid u, t) \\ &= 1 - e^{-ut}\end{aligned}$$

Estimating p from data

If we have 2 and 3 mutations on sister branches (same time to the MRCA) and a total of $M = 2000$ sites. For n branches with a vector y changes on each, and $k = \sum y$

$$\begin{aligned}
L(p) = \mathbb{P}(y_1 = 2, y_2 = 3 \mid p) &= \prod_{i=1}^{n} \binom{M}{y_i} p^{y_i}(1-p)^{M-y_i} \\
&= p^k (1-p)^{nM-k} \left(\prod_{i=1}^{n} \binom{M}{y_i}\right) \\
\ln L(p) &= k\ln(p) + (nM - k)\ln(1-p) + \sum_{i=1}^{n} \ln \binom{M}{y_i} \\
\frac{d \ln L(p)}{dp} &= \frac{k}{p} - \frac{nM - k}{1-p} \\
\frac{k}{\hat{p}} &= \frac{nM - k}{1 - \hat{p}} \\
k - k\hat{p} &= nM\hat{p} - k\hat{p} \\
\frac{k}{nM} &= \hat{p} = 5/4000 = 0.00125 \\
\ln L(p = 0.00125) &= 5\ln(p = 0.00125) + 3995\ln(0.99875) + \ln\left(\frac{2000 \times 1999}{2}\right) + \ln\left(\frac{2000 \times 1999 \times 1998}{6}\right) \\
&= -38.42 + \ln\left(\frac{2000 \times 1999}{2}\right) + \ln\left(\frac{2000 \times 1999 \times 1998}{2}\right) \\
&= -38.42 + 35.52 = -2.90
\end{aligned}$$

If we were testing against a null that $p_0 = 0.001$:

$$\begin{aligned}
\ln L(p = 0.001) &= 5\ln(p = 0.001) + 3995\ln(0.999) + 35.52 \\
&= -38.53 + 35.52 = -3.018
\end{aligned}$$

$$LRT = 2 \times (-2.90 + 3.018) = 2 \times 0.115 = 0.23$$

that is not close to our 3.84 cutoff which we find by looking up the $\chi^2_{df=1,\alpha=0.05}$ critical value.

conditioning on N instead of t

(this section was not a part of the homework, but I just wanted to illustrate that you can do the same sort of "integrate over all times if we want to learn about N" when we have the binomial likelihoods as we did with the Poisson. You get very similar estimators when the probability of change is rare.)

If we want to estimate the population size, we can integrate over t:

$$\begin{aligned}
\mathbb{P}(>0 \text{ mutations} \mid u, N) &= \int_0^\infty \left(1-e^{-ut}\right)\frac{1}{N}e^{-t/N}dt \\
&= \int_0^\infty \frac{1}{N}e^{-t/N} - \frac{1}{N}e^{-ut-t/N}dt \\
&= -e^{-t/N}\Big|_{t=0}^{t=\infty} - \frac{1}{N}\int_0^\infty e^{-\frac{(Nu+1)t}{N}}dt \\
&= -e^{-t/N}\Big|_{t=0}^{t=\infty} + \frac{1}{Nu+1}e^{-\frac{(Nu+1)t}{N}}\Big|_{t=0}^{t=\infty} \\
&= (0-(-1)) + \left(0 - \frac{1}{Nu+1}\right) \\
&= 1 - \frac{1}{Nu+1}
\end{aligned}$$

Then, we could do something like:

$$\begin{aligned}
1 - \frac{1}{\widehat{Nu}+1} &= \hat{p} \\
\frac{\widehat{Nu}}{\widehat{Nu}+1} &= \hat{p} \\
\widehat{Nu} &= \hat{p}\widehat{Nu} + \hat{p} \\
\widehat{Nu} - \hat{p}\widehat{Nu} &= \hat{p} \\
\widehat{Nu} &= \frac{\hat{p}}{1-\hat{p}} \\
&= \frac{k/nM}{1-k/nM} = \frac{k}{nM-k}
\end{aligned}$$

to convert our MLE of p to an MLE of Nu.