notes from the week of Feb. 11, 2019

Contents

1 Learning about a population's history from some sequence data

1.1 The data

We do some sequencing of mitochondrial DNA sequences and find that your sequence differs from JKK's sequence at 5 sites, so $X = 5$.

What can we learn from this?[1]

We might be able to learn something about:

- t the time since the most recent common ancestor (MRCA) of your mitochondrial genome and John's.
- u the rate at which mutations occur in the mitochondrial.
- (this one is less intuitive, but...) N the effective population size of humans.

[1] Ideally we'd think of a question, and then design the study and collect the data, but here we're thinking about the process of connecting a model to a dataset.

1.2 the likelihood

As always, the likelihood of our model is the probability of seeing a value of the random variable that represents our data set (X) which is equal to the data we actually saw (5) conditional on our model being true. We want $\mathbb{P}(X = 5 \mid \text{ some model})$, so we need to think about the model.

1.3 modeling mutations

We recall from intro bio that DNA polymerase is pretty good, which is to say that mutations are pretty rare. It also seems reasonable to think of them as not just rare, but fairly independent of each other. At least we don't think that if your mother had a mutation event in her mitochondrial genome that it probably has a large effect on the probability that you will experience a new mutational event.

So we could model mutations as a rare event that happens at some constant rate with the events being independent of each other. "Rare" is important because it lets us be somewhat confident that our observation of 5 *differences* between your sequence and JKK's implies that there were a total of 5 *mutations* that happened on the path that connects the MRCA genome to your mitochondrion and JKK's. So, for our purposes we'll interpret $X = 5$ differences to mean $X = 5$ mutations.

We don't really have good reason to think that the rate of mutations is constant, but it seems intractable to model multiple rates when we just have 1 datum, so let's just start with the one-rate model.

Statistical modeling is the art of browsing Wikipedia articles about probability distributions until you find one that makes (close to) the same assumptions that your biological model is making. In our case the opening paragraph[2] of https://en.wikipedia.org/wiki/Poisson_distribution tells us that when we have independent events occurring at some constant rate, then the Poisson distribution tells us the probability of any number of events.

Specifically if is the expected number of events, then:

$$\mathbb{P}(X = k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

That is nice. However, we were talking about rates and suddenly we've found a model that is parameterized in terms of the expected number of events. Recall (or see the Wikipedia article on expected value) that the expected value is basically the average value over infinitely many trials. We only have 1 trial, so how do we get the expected value?

If our process has constant rate u (mutations per generation) and runs for t generations from the MRCA to you and t generations from the MRCA to JKK, then we can see that:

$$\mathbb{E}(\#\text{ mutations}) = 2ut = \lambda \text{ (in the Poisson)} \tag{2}$$

where the funky $\mathbb{E}$ is a (nice) way to denote an expected value.

Now we can state a likelihood:

$$L(u,t) = \mathbb{P}(X = k \mid u,t) = \frac{(2ut)^k e^{-2ut}}{k!} \tag{3}$$

[2] At least on the Valentine's day, 2019 as I write this

where the feasible range of the parameters is $0 \leq u \leq \infty$ and $0 \leq t \leq \infty$. We could even fill in 5 for k to get

$$L(u,t) = \mathbb{P}(X = 5 \mid u,t) = \frac{1}{120}(2ut)^5 e^{-2ut} \tag{4}$$

If we chose, we could try to solve for the maximum liklihood estimates of u and t here. As we discussed in class, we'd run right into problems with non-identifiability (see https://en.wikipedia.org/wiki/Identifiability): u and t only occur in the likelihood equation in the product ut. So the data might prefer some value for ut, but we still wouldn't be able to learn about the value of u alone. In fact the non-identifiability is so extreme in this case that any value for u other than 0 will give us the same likelihood. So we can't learn *anything* that we didn't already know about u.

In this case it probably won't surprise you to learn that the MLE for the product $2ut$ is 5 (see the appendix A for a derivation). But we can get that by setting $u = 2.5$ and $t = 1$ or $u = 2.5^{-10}$ and $t = 10^{10}$ or any other pair of values on the hyperbola defined by $2ut = 5$.

1.4 Modeling population size

Aside from being able to astound friends with the fun fact about how many generations ago your maternal lineage shares and ancestor with JKK, it is not clear that you'd be all that interest in estimating t. If we'd sampled another couple of people, we'd (probably) get a different estimate for ut, and we often want to estimate parameters that have some generality about a part of nature.

So perhaps it is more sensible to ask whether or not we can learn something about the size of the human population from our data. To make (easy) progress we'd need to make some simplifying assumptions. The easiest model of formation of the next generation of individuals is just a model of non-overlapping generations with random reproduction. We want a model that connects the size of a population to a probability statement about time to the MRCA for two randomly selected gene copies.

1.4.1 intutition behind the coalescent

Clever population geneticists realized that it was easier to make this model if we look backward and think about the question: "if I've selected your mitochondrial genome and JKK's, then what is the probability that the first generation that those mitochondria come from the same woman was generation g?" We recall from high school biology that the mitochondrion is (1) the powerhouse of the cell, and (2) inherited maternally.[3] That is why we are just thinking about sharing a female ancestor.

The most naive model just says that if there are N potential mom's in a generation, then the chance that two randomly chosen individuals share a mom in the previous generation is simply $1/N$. So,

$$\mathbb{P}(g = 1 \mid N) = \frac{1}{N} \text{ and so } \mathbb{P}(g > 1 \mid N) = 1 - \frac{1}{N}$$

The jargon for "sharing an ancestor" is called a coalescence event. If the population size is constant (and reproduction is random) then the probability of coalescence of two randomly selected gene copies in each generation is $1/N$ and the probability of failure to coalesce in that generation is $1 - \frac{1}{N}$.

[3] which is not quite true: https://www.ncbi.nlm.nih.gov/pubmed/12192017

So, what is the chance that the most recent shared member of the maternal lineage is 2 generations ago? Well that means "no shared mom, but a shared grandmom" so:

$$\mathbb{P}(g = 2 \mid N) = \left(1 - \frac{1}{N}\right)\left(\frac{1}{N}\right)$$

In general the form is

$$\mathbb{P}(G = g \mid N) = \left(1 - \frac{1}{N}\right)^{g-1}\left(\frac{1}{N}\right)$$

(which our modeling browsing tells us is the beautiful geometric distribution).

1.4.2 the continuous-time version of the coalescent that we actually use

It turns out that even if we use a more realistic model of reproduction (with overlapping generations and realistic numbers of offspring/parent, *etc.*) that we can get decent approximations of the time to common ancestor in our situation using the probability density:

$$f(t \mid N) \quad = \quad \left(\frac{1}{N}\right) e^{-\frac{t}{N}} \tag{5}$$

where t is the generation expressed as a continuous variable. This is the (even more beautiful) exponential distribution. As wikipedia notes, it is the "continuous analogue" of the geometric distribution. It gives a probability density function for the waiting time for the first event to occur when the events occur stochastically but at some constant rate (here $1/N$) while the geometric gives us a probability distribution on the count of how many trials before the first success (when you have repeated trials with the same probability of success on each trial).

Both distributions give waiting times. It is easier to think about discrete variables (and it might make more sense to think about a discrete number of generations). Nevertheless, it is going to be easier to use the exponential. In the next step, we are going to try to think about modeling how many mutations we'd see when drawing from a population when we don't know the time to MRCA. That will entail summing over all possible times from 0 to infinity. An integration is usually easier than infinite sums. So we frequently use continuous approximations of processes that we might occur to as as being discrete.

1.4.3 Connecting the population size to our data

We started this section by noting that we might not care about t, but be more intested in N. But if we want to learn about N we need it to be in the likelihood equation.

As JKK pointed out, we frequently use the law of total probability to calculate the probability of an event when we can figure out how to express it's probability only in terms of some additional information.

In our case we had a likelihood in terms of u and t, but we want to focus on u and N instead. Our brief interlude into coalescent theory gave us some insight into the waiting time to the MRCA, but we always need to make a probability statement about the form that our data takes when we use likelihood. Fortunately, we know about $\mathbb{P}(X = k \mid u, t)$ and we know about the probability distribution of waiting times t to the MRCA, so we can combine them.

The law of total probability (when conditioning on a real, continuous variable is):

$$\mathbb{P}(A) = \int_{-\infty}^{\infty} \mathbb{P}(A \mid B) f(B) dB \tag{6}$$

Applying this to our problem:

$$\begin{aligned} L(u,N) &= \int_0^{\infty} \mathbb{P}(X = k \mid u, t) f(t \mid N) dt && (7) \\ &= \int_0^{\infty} \left(\frac{(2ut)^k e^{-2ut}}{k!}\right)\left(\frac{1}{N}\right) e^{-t/N} dt && (8) \end{aligned}$$

where we've set the lower bound of the integration to be 0 because time to MRCA can't be negative. This is where JKK encouraged you to use Wolfram Alpha (https://www.wolframalpha.com/) for help in integrating. That is a good idea (but do take a look at appendix B).

We find that:

$$L(u,N) = \frac{(2uN)^k}{(2uN+1)^{k+1}} \tag{9}$$

Once again we see that our two parameters are entangled and we have a case of non-identifiability. Whenever N shows up in the likelihood it is being multiplied by u, so we can't hope to have the data separate the 2 parameters.

Population geneticists throw up their hands at this point and settle for estimating a mutation-rate scaled population size:

$$\theta = 2uN.$$

Making that substitution and going through our recipe for finding the maximum likelihood estimator:

$$\begin{aligned} L(\theta) &= \frac{\theta^k}{(\theta+1)^{k+1}} && (10) \\ \ln L(\theta) &= k \ln\theta - (k+1)\ln(\theta+1) && (11) \\ \frac{d \ln L(\theta)}{d\theta} &= \frac{k}{\theta} - \frac{k+1}{\theta+1} && (12) \\ \frac{k}{\hat{\theta}} - \frac{k+1}{\hat{\theta}+1} &= 0 && (13) \\ \frac{k}{\hat{\theta}} &= \frac{k+1}{\hat{\theta}+1} && (14) \\ k\hat{\theta} + k &= k\hat{\theta} + \hat{\theta} && (15) \\ k &= \hat{\theta} && (16) \end{aligned}$$

Once again, we find that the MLE is sort of the obvious answer, but it is still useful to work in the likelihood framework. We could calculate the log-likelihood at the MLE and then find the values of θ that have $\ln L$ values of 1.92 lower than the maximum to get a nice interval estimate.

2 Extending to richer data

What if we knew that 2 changes occurred along the line from the MRCA to you and 3 occurred on the lineage leading to JKK's mitochondrion?

Would this change our estimate of the population size? Let's call this representation of the data Y to keep from getting confused with the previous example. So previously the data was $X = 5$, and now the data is the vector $Y = [y_1 = 2, y_2 = 3]$.

Assuming evolution on the 2 branches of the genealogy is independent:

$$\mathbb{P}(Y \mid u, N) \quad = \quad \prod_{i=1}^{n} \mathbb{P}(y_i \mid u, N) \tag{17}$$

where n is the sample size (2 when we have 2 branches). Conceptually the model is the same but the expected number of substitutions is just ut along each branch instead of $2ut$ across the pair branches.

Making the new version of Equation (8)

$$\begin{aligned} L(u,N) = \mathbb{P}(Y \mid u, N) &= \int_0^\infty \mathbb{P}(Y \mid u, t) f(t \mid N) dt \qquad (18)\\ &= \int_0^\infty \left[\prod_{i=1}^{n} \mathbb{P}(y_i \mid u, N)\right] f(t \mid N) dt \qquad (19)\\ &= \int_0^\infty \left[\prod_{i=1}^{n} \left(\frac{(ut)^{y_i} e^{-ut}}{y_i!}\right)\right] \left(\frac{1}{N}\right) e^{-t/N} dt \qquad (20) \end{aligned}$$

At this point, we can work on simplifiying the product that is in the hard braces:

$$\begin{aligned} \prod_{i=1}^{n} \left(\frac{(ut)^{y_i} e^{-ut}}{y_i!}\right) &= \frac{\prod_{i=1}^{n} \left[(ut)^{y_i} e^{-ut}\right]}{\prod_{i=1}^{n} y_i!} \qquad (21)\\ &= \frac{\left[(ut)^{\sum_{i=1}^{n} y_i}\right] \left[e^{-nut}\right]}{\prod_{i=1}^{n} y_i!} \qquad (22) \end{aligned}$$

Recall that previously we were just thinking of the sum of number of counts across each branch ($X = k = 5$ instead of $y_i = 2, y_2 = 3$). Consider the simplification in Equation 22 with $n = 2$ and $y_i = 2, y_2 = 3$: :

$$\frac{(ut)^{2+3} e^{-2ut}}{2!3!} = \frac{(ut)^5 e^{-2ut}}{2!3!} = \left(\frac{1}{2!3!}\right) (ut)^5 e^{-2ut}$$

to the corresponding part of equation 8 when the total count of mutations is $k = 5$:

$$\frac{(2ut)^5 e^{-2ut}}{5!} = \left(\frac{2^5}{5!}\right) (ut)^5 e^{-2ut}$$

We see that the likelihoods under these two models only differ by their constant coefficients. The coefficients are $\frac{1}{\prod_{i=1}^{n} y_i!}$ and $\frac{2^k}{k_i!}$ where $k = \sum_{i=1}^{n} y_i$. Those expressions contain variables to the data, so they may not appear to be "constant," but you have to recall that when we are doing inference the data is fixed. As you consider different values of the parameters u and N those coefficients don't change.

Whenever you calculate the likelihood ratio of two different values of a parameter, the coefficients will cross out in the ratio. In other words, the expression of the likelihood tells us that we will not get any more information about u and N from the data with $Y = [2, 3]$ than we did from $X = 5$. The MLE and the interval estimates will be the same as they were in the previous section.

3 Extending to a richer model

Can we tell if our data prefers a model that allows the rate of mutation to differ on the lineage leading to you vs the lineage leading to JKK?

Now we have to split our rate from u into a richer model with two rate parameters, u_1 and u_2. The modification to the likelihood is straightforward: now the parameters get subscripts, too!

$$L(u_1, u_2, N) = \mathbb{P}(Y \mid u_1, u_2, N) \quad = \quad \int_0^\infty \left[\prod_{i=1}^{n} \left(\frac{(u_i t)^{y_i} e^{-u_i t}}{y_i!}\right)\right] \left(\frac{1}{N}\right) e^{-t/N} dt \tag{23}$$

Appendix C goes through the gory details.

You end up with 2 estimators for $\hat{\theta_1} = y_1$ and $\hat{\theta_2} = y_2$ just as you would hope/expect. Crucially if you want to test the one-rate model vs the two rate-model, you have to use the equation (23) for your likelihood calculation.

The likelihood calculated on based on the sum (and shown in equation 9) has different constants from the one in Equation 23, because the two forms of the data $X = k = 5$ vs $Y = [y_1 = 2, y_2 = 3]$ are slightly different. We can only compare likelihoods when calculated on the same representation of the data!

A Deriving the MLE of the product ut

We noted the fact that ut only occur as a product, so lets work with the product and call it $\nu = 2ut$. Recall the recipe is to find the derivative with respect to the parameter(s), and find the parameter value that makes that derivative 0, and that we usually find it easier to do this on the log scale. We'll start with Equation (4):

$$\begin{aligned}
L(u,t) = \mathbb{P}(X = 5 \mid u, t) &= \frac{1}{120}(2ut)^5 e^{-2ut} \\
&= \frac{1}{120}(\nu)^5 e^{-\nu} \\
\ln L(\nu) &= \ln(1/120) + 5\ln\nu - \nu \\
\frac{d \ln L(\nu)}{d\nu} &= \frac{5}{\nu} - 1 \\
\frac{5}{\hat{\nu}} - 1 &= 0 \\
5 &= \hat{\nu}
\end{aligned}$$

B The tricky integral

Equation (8) said:

$$\mathbb{P}(X = k \mid u, N) = \int_0^\infty \left(\frac{(2ut)^k e^{-2ut}}{k!}\right)\left(\frac{1}{N}\right) e^{-t/N} dt$$

I'm not going to derive this, as it is beyond me. But I'll sketch how you can make progress. First note that (with respect to the integration, not the larger likelihood problem) u and k and N are all constants. Integration is usually easier if we move the constants out front. But we can only do that for coefficients that are entirely constant (don't have t in our case, becase we are integrating over dt), so:

$$\begin{aligned}
\int_0^\infty \left(\frac{(2ut)^k e^{-2ut}}{k!}\right)\left(\frac{1}{N}\right) e^{-t/N} dt &= \int_0^\infty \left(\frac{(2u)^k}{k!}\right) t^k e^{-2ut} \left(\frac{1}{N}\right) e^{-t/N} dt \\
&= \int_0^\infty \left(\frac{(2u)^k}{N(k!)}\right) t^k e^{-2ut} e^{-t/N} dt \\
&= \left(\frac{(2u)^k}{N(k!)}\right) \int_0^\infty t^k e^{-2ut-(t/N)} dt \\
&= \left(\frac{(2u)^k}{N(k!)}\right) \int_0^\infty t^k e^{-\frac{2uN+1}{N}t} dt \\
&= \left(\frac{(2u)^k}{N(k!)}\right) \int_0^\infty t^k e^{-Ct} dt
\end{aligned}$$

where $C = \frac{2uN+1}{N}$. It'll be easier to look up $\int_0^\infty t^k e^{-Ct} dt$ where k is a non-negative integer and C is a positive real number using Wolfram Alpha or some book than it would be to search for something that looks like Equation (8).

In fact Wolfram Alpha give you the result shown at: https://www.wolframalpha.com/input/?i=integrate+(x%5Ek+exp%5B-C+x%5D)+dx+from+0+to+infinity

The part of the solution that says "$\mathrm{Re}(k) > -1 \wedge \mathrm{Re}(C) > 0$" means "when k is a real number greater than -1 and C is a positive real number" fortunately for us, both of those hold for our problem.[4] So the solution is what is shown prior to that qualification about the domain, namely:

$$\int_0^\infty t^k e^{-Ct} dt = C^{-k-1}\Gamma(k+1)$$

where $\Gamma(k+1)$ is the Gamma function. For a positive integer argument m it is an offset factorial $\Gamma(m) = (m-1)!$. So in our solution $\Gamma(k+1) = k!$.

Substituting in for C and remembering the constant coefficient for the integral we have:

$$\begin{aligned}
\mathbb{P}(X = k \mid u, N) &= \left(\frac{(2u)^k}{N(k!)}\right) \int_0^\infty t^k e^{-Ct} dt \\
&= \left(\frac{(2u)^k}{N(k!)}\right) C^{-k-1}\Gamma(k+1) \\
&= \left(\frac{(2u)^k}{N(k!)}\right) \left(\frac{2uN+1}{N}\right)^{-k-1} (k!) \\
&= \left(\frac{(2u)^k}{N(k!)}\right) \left(\frac{N}{2uN+1}\right)^{k+1} (k!) \\
&= \left(\frac{(2u)^k}{N}\right) \left(\frac{N^{k+1}}{(2uN+1)^{k+1}}\right) \\
&= \frac{(2u)^k N^k}{(2uN+1)^{k+1}} \\
&= \frac{(2uN)^k}{(2uN+1)^{k+1}}
\end{aligned}$$

C The two-mutation-rate tricky integral

We'll use $w = \sum_{i=1}^n u_i$ and $k = \sum i = 1^n y_i$ here.

$$\begin{aligned}
\int_0^\infty \left[\prod_{i=1}^n \left(\frac{(u_i t)^{y_i} e^{-u_i t}}{y_i!}\right)\right] \left(\frac{1}{N}\right) e^{-t/N} dt &= \int_0^\infty \left(\frac{\prod_{i=1}^n \left[(u_i t)^{y_i} e^{-u_i t}\right]}{\prod_{i=1}^n y_i!}\right) \left(\frac{1}{N}\right) e^{-t/N} dt && (24) \\
&= \int_0^\infty \left(\frac{e^{-wt} t^k \prod_{i=1}^n \left[u_i^{y_i}\right]}{\prod_{i=1}^n y_i!}\right) \left(\frac{1}{N}\right) e^{-t/N} dt && (25) \\
&= \frac{\prod_{i=1}^n \left[u_i^{y_i}\right]}{N \prod_{i=1}^n y_i!} \int_0^\infty \left(e^{-wt} t^k\right) e^{-t/N} dt && (26) \\
&= \frac{\prod_{i=1}^n \left[u_i^{y_i}\right]}{N \prod_{i=1}^n y_i!} \int_0^\infty \left(e^{-\frac{Nw+1}{N} t} t^k\right) dt && (27) \\
&= \frac{\prod_{i=1}^n \left[u_i^{y_i}\right]}{N \prod_{i=1}^n y_i!} \int_0^\infty t^k e^{-Dt} dt && (28)
\end{aligned}$$

[4] https://en.wikipedia.org/wiki/List_of_mathematical_symbols is quite handy if you encounter math symbols that you don't understand

where $D = \frac{Nw+1}{N}$ which is the same integral as in appendix 8, but with a different constant (and a different coefficient outside of the integration).

$\mathbb{P}(Y = [y_1, y_2] \mid u_1, u_2, N)$ is equal to:

$$\begin{aligned}
& \frac{\prod_{i=1}^{n}[u_i^{y_i}]}{N\prod_{i=1}^{n} y_i!} \int_0^\infty t^k e^{-Dt} dt \\
= & \frac{\prod_{i=1}^{n}[u_i^{y_i}]}{N\prod_{i=1}^{n} y_i!} \left(\frac{Nw+1}{N}\right)^{-k-1} \Gamma(k+1) \\
= & \frac{\prod_{i=1}^{n}[u_i^{y_i}]}{N\prod_{i=1}^{n} y_i!} \left(\frac{N}{Nw+1}\right)^{k+1} k! \\
= & \left[\prod_{i=1}^{n}\left(\frac{u_i^{y_i}}{y_i!}\right)\right]\left(\frac{N^k}{(Nw+1)^{k+1}}\right) k!
\end{aligned}$$

Moving to the log scale, and collapsing all terms that don't depend on u_1, u_2 or N into constant G, and expanding w to $u_1 + u_2$, we get:

$$\begin{aligned}
\ln L &= G + \left(\sum_{i=1}^{n} y_i \ln(u_i)\right) + k\ln N - (k+1)\ln(Nu_1 + Nu_2 + 1) \\
\frac{d\ln L}{du_1} &= \frac{y_1}{u_1} - \frac{(k+1)N}{Nu_1 + Nu_2 + 1} \\
\frac{y_1}{\hat{u_1}} & - = \frac{(k+1)N}{N\hat{u_1} + Nu_2 + 1} \\
y_1 N\hat{u_1} + y_1 N u_2 + y_1 &= kN\hat{u_1} + N\hat{u_1}
\end{aligned}$$

recalling that $k = y_1 + y_2$:

$$\begin{aligned}
y_1 N u_2 + y_1 &= kN\hat{u_1} + N\hat{u_1} - y_1 N\hat{u_1} \\
y_1(Nu_2 + 1) &= (y_2 + 1)N\hat{u_1} \\
\hat{u_1} &= \frac{y_1(Nu_2 + 1)}{(y_2 + 1)N}
\end{aligned}$$

Similarly:

$$\hat{u_2} = \frac{y_2(Nu_1 + 1)}{(y_1 + 1)N}$$

The global maximum likelihood estimator is when both u_1 and u_2 are maximized. So we can swap in $\hat{u_2}$ for u_2 in the equation for $\hat{u_1}$:

$$\begin{aligned}
\hat{u_1} &= \frac{y_1\left(N\frac{y_2(N\hat{u_1}+1)}{(y_1+1)N} + 1\right)}{(y_2+1)N} \\
(y_2+1)N\hat{u_1} &= y_1\left(N\frac{y_2(N\hat{u_1}+1)}{(y_1+1)N} + 1\right) \\
&= y_1\left(\frac{y_2(N\hat{u_1}+1)}{(y_1+1)} + 1\right)
\end{aligned}$$

$$\begin{aligned}
&= y_1\left(\frac{y_2(N\hat{u_1}+1)+y_1+1}{(y_1+1)}\right) \\
(y_1+1)(y_2+1)N\hat{u_1} &= (y_2(N\hat{u_1}+1)+y_1+1)y_1 \\
y_1y_2N\hat{u_1}+(y_{1+}y_2+1)N\hat{u_1} &= y_1y_2N\hat{u_1}+(y_2+y_1+1)y_1 \\
(y_{1+}y_2+1)N\hat{u_1} &= (y_2+y_1+1)y_1 \\
\hat{u_1} &= \frac{y_1}{N}
\end{aligned}$$

In otherwords our best estimate of Nu_1 is y_1, just as we would have guessed.