notes from Monday, Feb. 4, 2019

Contents

1 Estimating allele freq from genotypes of mom + one child families

1.1 The data

n families that have a genotype of the mother and one child. For each family we have a pair of observations, a mother's genotype and a child's genotype. So $x_i = (m_i, c_i)$ if x_i is the pair of genotypes for family i with m_i and c_i being the mother's genotype and child's genotype.

1.2 model

We have 1 parameter that we want to estimate: q, the frequency of the A allele (with B being the other allele). So:

$$\mathbb{P}(g = A) = q \quad (1)$$

$$\mathbb{P}(g = B) = 1 - q \quad (2)$$

where g represents some gene sequence drawn randomly from the "gene pool."

1.3 The likelihood

By definition the likelihood is the probability of the entire data set: $L(q) = \mathbb{P}(X \mid q)$ Assuming the families are independent:

$$L(q) = \mathbb{P}(X \mid q) \quad = \quad \prod_{i=1}^{n} \mathbb{P}(x_i) \tag{3}$$

It is tempting to simplify by assuming the genotypes are independent:

$$L(q) \quad = \quad \prod_{i=1}^{n} \mathbb{P}(m_i)\mathbb{P}(c_i), \text{ but } \textbf{this is wrong!}$$

This is wrong because it assumes that the child's genotype is independent of the mom's genotype. That is clearly not a reasonable assumption. We need to use conditional probability:

$$L(q) \quad = \quad \prod_{i=1}^{n} \left[\mathbb{P}\left(m_i\right) \mathbb{P}\left(c_i \mid m_i\right)\right] \tag{4}$$

1.4 Connecting the model to the general likelihood equation

1.4.1 The probability of maternal genotypes

The first factor for each family is the probability that we'd get mom's genotype, if the allele frequency of A was q. If we assume mom's two gene copies are each independent draws from the "gene pool" we can make probabilistic statements about mom's genotype from the allele frequencies (this is just the Hardy-Weinberg model).

$$\mathbb{P}(m_i = AA) \quad = \quad q^2 \tag{5}$$
$$\mathbb{P}(m_i = AB) \quad = \quad 2(1-q)q \tag{6}$$
$$\mathbb{P}(m_i = BB) \quad = \quad (1-q)^2 \tag{7}$$

1.4.2 The probability of the child's genotype given mom's genotype

The second factor is the probability of each possible value for c_i conditional on m_i and q. Mendel tells us these. It is convenient to put them in table 1:

	$c_i = AA$	$c_i = AB$	$c_i = BB$
$m_i = AA$	q	$(1-q)$	0
$m_i = AB$	$q/2$	$1/2$	$(1-q)/2$
$m_i = BB$	0	q	$(1-q)$

Table 1: $\mathbb{P}(c_i \mid m_i, q)$ for every combiniation of m_i and c_i

The middle row is a bit trickier to derive than the others:

$$\begin{aligned}
\mathbb{P}(c_i = AA \mid m_i = AB) &= \mathbb{P}(\text{mom gives } A)\mathbb{P}(\text{dad gives } A) \\
&= (1/2)(q) = q/2 \qquad (8) \\
\mathbb{P}(c_i = AB \mid m_i = AB) &= \mathbb{P}(\text{mom gives } A)\mathbb{P}(\text{dad gives } B) + \mathbb{P}(\text{mom gives } B)\mathbb{P}(\text{dad gives } A) \\
&= (1/2)(1-q) + (1/2)(q) = 1/2 \qquad (9)
\end{aligned}$$

Note that each row of the conditional probability table sums to 1, because each row considers the entire sample space of the random event c_i.

1.5 Expressing the likelihood as the joint probability for each "row" of data

Making the probability statement for each family, is then just the product of the probability of mom's genotype and the probability of the child conditional on mom's genotype (the $\mathbb{P}(m_i)\,\mathbb{P}(c_i \mid m_i)$ factors in the likelihood eqn).

	$c_i = AA$	$c_i = AB$	$c_i = BB$
$m_i = AA$	q^3	$q^2(1-q)$	0
$m_i = AB$	$q^2(1-q)$	$q(1-q)$	$q(1-q)^2$
$m_i = BB$	0	$q(1-q)^2$	$(1-q)^3$

Table 2: $\mathbb{P}(m_i, c_i \mid q)$ for all combinations of m_i and c_i

It is a bit harder to see, but the sum of every entry in table 3 is to 1 (for any value of q in the feasible range $0 \leq q \leq 1$), because this table covers all of the possible genotypes for a mom+child.

1.6 Terser representation of the data

We don't care what order we encounter the families. The only thing that matters is the number of families observed in each datatype. Also note that geneticist like to be terse. So we might want to denote a genotype with a count of B alleles (so $AA = 0$, $AB = 1$, $BB = 2$). This let's us summarize the data in nine counts: $[n_{00}, n_{01}, n_{02}, n_{10}, \ldots, n_{22}]$ where

$$n = \sum_{a=0}^{2}\sum_{b=0}^{2} n_{ab}$$

or in tabular form:

Note that if $n_{02} > 0$ or $n_{20} > 0$, we have an "impossible family" (indicating an error in our data or a mutation, which is disallowed in our model).

Because the order of the observations don't matter (only the counts matter), the stats jargon would say the the counts are a "sufficient statistic" in the inference problem.

	$c_i = AA$	$c_i = AB$	$c_i = BB$	row sum
$m_i = AA$	n_{00}	n_{01}	n_{02}	n_0
$m_i = AB$	n_{10}	n_{11}	n_{12}	n_1
$m_i = BB$	n_{20}	n_{21}	n_{22}	n_2
Grand total				n

Table 3: conventions for referring to counts of the data

1.7 Rewriting the likelihood with the terser data representation

This make the likelihood contribution very compact, too:

$$\begin{aligned} L(q) &= \prod_{i=1}^{n} \left[\mathbb{P}\left(m_i\right) \mathbb{P}\left(c_i \mid m_i\right)\right] \\ &= \prod_{a=0}^{2} \prod_{b=0}^{2} \left[\mathbb{P}\left(m_i = a\right) \mathbb{P}\left(c_i = b \mid m_i = a\right)\right]^{n_{ab}} \end{aligned} \tag{10}$$

1.8 Moving to the log-scale

Products are hard to deal with when finding derivatives, so we can move to the log scale:

$$\begin{aligned} \ln L(q) &= \sum_{a=0}^{2} \sum_{b=0}^{2} \ln\left(\left[\mathbb{P}\left(m_i = a\right) \mathbb{P}\left(c_i = b \mid m_i = a\right)\right]^{n_{ab}}\right) \\ &= \sum_{a=0}^{2} \sum_{b=0}^{2} n_{ab} \ln\left(\left[\mathbb{P}\left(m_i = a\right) \mathbb{P}\left(c_i = b \mid m_i = a\right)\right]\right) \qquad (11) \\ &= \sum_{a=0}^{2} \sum_{b=0}^{2} \left(n_{ab} \ln\left[\mathbb{P}\left(m_i = a\right)\right] + n_{ab} \ln\left[\mathbb{P}\left(c_i = b \mid m_i = a\right)\right]\right) \qquad (12) \end{aligned}$$

Note that there is some sharing of terms in the log-likelihood, if we just look at the first term in each log-likelihood:

$$\begin{aligned} \ln L(q) &= \sum_{a=0}^{2} \sum_{b=0}^{2} n_{ab} \ln\left[\mathbb{P}\left(m_i = a\right)\right] + \sum_{a=0}^{2} \sum_{b=0}^{2} n_{ab} \ln\left[\mathbb{P}\left(c_i = b \mid m_i = a\right)\right] \qquad (13) \\ &= \sum_{a=0}^{2} n_a \ln\left[\mathbb{P}\left(m_i = a\right)\right] + \sum_{a=0}^{2} \sum_{b=0}^{2} n_{ab} \ln\left[\mathbb{P}\left(c_i = b \mid m_i = a\right)\right] \qquad (14) \end{aligned}$$

where we use the row sum notation introduced earlier: $n_a = n_{a0} + n_{a1} + n_{a2}$ to denote a count of all of the families that have a particular maternal genotype. In other words, if you just see one number as a subscript for n we are using it to refer to the count of mothers with that genotype. The notation allows us to skip the summation over b (our index indicating the child's genotype).

This lets us realize that the log-likelihood neatly breaks down into a term that reflects the probability of the maternal genotypes, and another term that reflects the probability of the children's genotypes:

$$\ln L(q) = M + C \text{ where:}$$

$$
\begin{aligned}
M &= \sum_{a=0}^{2} n_a \ln\left[\mathbb{P}\left(m_i = a\right)\right] \\
C &= \sum_{a=0}^{2}\sum_{b=0}^{2} n_{ab} \ln\left[\mathbb{P}\left(c_i = b \mid m_i = a\right)\right]
\end{aligned}
$$

1.9 Making the log-likelihood equation more concrete

Now we can substitute our probability model into the likelihood. The algebra in parts:

$$
\begin{aligned}
M &= n_0 \ln(q^2) + n_1 \ln(2q(1-q)) + n_2 \ln((1-q)^2) \\
&= 2n_0 \ln(q) + n_1 \ln(2) + n_1 \ln(q) + n_1 \ln(1-q)) + 2n_2 \ln(1-q) \\
&= (2n_0 + n_1)\ln(q) + (2n_2 + n_1)\ln(1-q) + n_1 \ln(2) \qquad (15)
\end{aligned}
$$

There is less simplification when we substitute for C. Dropping out the terms with coefficients (n_{02} and n_{20}) we see:

$$
\begin{aligned}
C &= n_{00}\ln(q) + n_{01}\ln(1-q) + n_{10}\ln(q/2) + n_{11}\ln(1/2) + n_{12}\ln((1-q)/2) + n_{21}\ln(q) + n_{22}\ln(1-q) \\
&= (n_{00} + n_{10} + n_{21})\ln(q) + (n_{01} + n_{12} + n_{22})\ln(1-q) + (n_{10} + n_{11} + n_{12}+)\ln(1/2) \\
&= (n_{00} + n_{10} + n_{21})\ln(q) + (n_{01} + n_{12} + n_{22})\ln(1-q) - n_1 \ln(2) \qquad (16)
\end{aligned}
$$

So the full log-likelihood collapses to a charming form:

$$
\begin{aligned}
\ln L(q) &= M + C \\
&= (2n_0 + n_1)\ln(q) + (2n_2 + n_1)\ln(1-q) + n_1\ln(2)\ldots \\
&\quad + (n_{00} + n_{10} + n_{21})\ln(q) + (n_{01} + n_{12} + n_{22})\ln(1-q) - n_1\ln(2) \\
&= (2n_0 + n_1)\ln(q) + (2n_2 + n_1)\ln(1-q) + (n_{00} + n_{10} + n_{21})\ln(q) + (n_{01} + n_{12} + n_{22})\ln(1-q) \\
&= (2n_0 + n_1 + n_{00} + n_{10} + n_{21})\ln(q) + (2n_2 + n_1 + n_{01} + n_{12} + n_{22})\ln(1-q) \qquad (17)
\end{aligned}
$$

1.10 Finding the maximum likelihood estimator of q

We now will differentiate the log-likelihood with respect to the parameter of the model (q), so we can see how the log-likelihood changes as a function of q:

$$
\begin{aligned}
\ln L(q) &= (2n_0 + n_1 + n_{00} + n_{10} + n_{21})\ln(q) + (2n_2 + n_1 + n_{01} + n_{12} + n_{22})\ln(1-q) \\
\frac{d\ln L(q)}{dq} &= \frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{q} - \frac{2n_2 + n_1 + n_{01} + n_{12} + n_{22}}{1-q} \qquad (18)
\end{aligned}
$$

The maximum for the log-likelihood has to be either at one of the ends of the feasible range (at $q = 0$ or $q = 1$) or when the derivative is 0. Assuming (for now) that the MLE is where the derivative is 0, we can solve for the MLE ($\hat{q}$):

$$
\frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{\hat{q}} - \frac{2n_2 + n_1 + n_{01} + n_{12} + n_{22}}{1-\hat{q}} = 0
$$

$$
\begin{aligned}
\frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{\hat{q}} &= \frac{2n_2 + n_1 + n_{01} + n_{12} + n_{22}}{1 - \hat{q}} \\
(2n_0 + n_1 + n_{00} + n_{10} + n_{21})(1 - \hat{q}) &= (2n_2 + n_1 + n_{01} + n_{12} + n_{22})\hat{q} \\
(2n_0 + n_1 + n_{00} + n_{10} + n_{21}) &= (2n_0 + n_1 + n_{00} + n_{10} + n_{21} + 2n_2 + n_1 + n_{01} + n_{12} + n_{22})\hat{q} \\
(2n_0 + n_1 + n_{00} + n_{10} + n_{21}) &= (2n_0 + 2n_1 + 2n_2 + n_{00} + n_{10} + n_{21} + n_{01} + n_{12} + n_{22})\hat{q} \\
\hat{q} &= \frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{2n_0 + 2n_1 + 2n_2 + n_{00} + n_{10} + n_{21} + n_{01} + n_{12} + n_{22}}
\end{aligned}
$$

Note that $n_0 = n_{00} + n_{01}$ because $n_{02} = 0$ if there are no impossible families. Similarly: $n_2 = n_{21} + n_{22}$ This lets us simplify the denominator:

$$
\hat{q} = \frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{3n_0 + 2n_1 + 3n_2 + n_{10} + n_{12}}
$$

A bit more cryptically: $n_1 = n_{10} + n_{11} + n_{12}$, so $n_{10} + n_{12} = n_1 - + n_{11}$, allowing us to restate:

$$
\hat{q} = \frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{3n_0 + 3n_1 + 3n_2 - n_{11}}
$$

Further $n = n_0 + n_1 + n_2$, so

$$
\hat{q} = \frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{3n - n_{11}}
$$

Making sense of the MLE

When we observe n unrelated, diploid mom's drawn from the population, we would expect that the best guess of the frequency of A alleles is the number of A alleles divided by the number of loci that we have sequenced:

$$
\hat{q}_m = \frac{2n_0 + n_1}{2n} \tag{19}
$$

where the coefficient of 2 in the numerator comes from the fact that n_0 mom's have 2 A alleles. The 2 in the denominator reflect diploidy (2 gene copies/mom).

If we know mom's genotype, then the only thing we learn about the frequency of alleles is what we learn by looking at dad's contribution. If we denote dad's contribution by d_i we can line up the outcomes in our tabular form:

	$c_i = AA$	$c_i = AB$	$c_i = BB$
$m_i = AA$	$d_i = A$	$d_i = B$	impossible
$m_i = AA$	$d_i = A$	$d_i =?$	$d_i = B$
$m_i = AA$	impossible	$d_i = A$	$d_i = B$

You might expect an estimator based soley on the children's genotypes to be just a count of all of the times dad gave an A over all of the times that we can figure out what dad gave. This turns out to be:

$$
\hat{q}_c = \frac{n_{01} + n_{10} + n_{21}}{n - n_{11}} \tag{20}
$$

because the only case in which we can't figure out dad's contribution is n_{11} families (those for which mom and child have genotype AB).

Seen in light of these facts our overall estimator:

$$\hat{q} \quad = \quad \frac{2n_0 + n_1 + n_{00} + n_{10} + n_{21}}{3n - n_{11}}$$

makes sense because the numerator is the number of counts in of A alleles that can be clearly seen to be drawn from the gene pool (the sum of the numerators of the $\hat{q}_m$ and $\hat{q}_c$ estimators). And the denominator is just the count of the scorable-as-drawn-from-the-gene-pool alleles (the sum of denominators of the 2 partial-data estimators).

MLE's don't always have such a nice intuitive interpretation to them, so we must cherish the cases in which the do.