Bayesian estimation

Kent E. Holsinger

25 July 2008

Imagine the following situation. We've determined the blood type of a young girl and her mother at the M/N locus. The girl's blood type is MN. Her mother's blood type is N. Since the alleles at this locus are co-dominant, we know that the mother is homozygous for the N locus. But what about the father? We know that he passed an M allele to his daughter, but was he homozygous for the M allele or an MN heterozygote? We could use Mendel's rules and the principle of maximum likelihood to answer this question.

If the the father was homozygous for M, the probability that he passed the M allele to his daughter is 1 (ignoring mutation). If he was heterozygous, the probability is 1/2. Let's let X be a random variable representing the allele that the father passed along to his daughter, and let's let G be a random variable representing the (unknown) genotype of the father. In our case, $X = M$ and G is either MM or MN. The maximimum likelihood estimate of the father's genotype is the value of G that maximizes $\mathrm{P}(X|G)$. Since $1 > 1/2$, the maximum likelihood estimate of G is MM. In other words, we'd guess that the father was homozygous for M.

But suppose we knew that in this population the M allele has a frequency of only 1%. Given that this is a human population, genotypes are likely to be in Hardy-Weinberg proportions, so the frequency of the homozygous genotype is only $0.01 \times 0.01 = 0.0001$, or 0.01%, while the frequency of the heterzygous genotype is $2 \times 0.01 \times 0.99 = 0.0198$, or almost 2%. Even though the principle of maximum likelihood would tell us that the father's is homozygous for M, the frequency of that genotype is so small, it seems much more likely that the father was a heterozygote. But how do we take into account this new information about the allele frequency in the population? First, we have to take a little digression into probability theory.

Bayes' theorem

In the example we just discussed, we wrote the likelihood of the data as $\mathrm{P}(X|G)$, where X is the random variable corresponding to our data (the identity of the allele the girl inherited from her father) and G is the (unknown) genotype of her father. The probability $\mathrm{P}(X|G)$ is an example of what statisticians call a conditional probability. It's the probability that event X happens – the girl inherits the M allele from her father – given that we already know that the father has a particular genotype – either MM or MN. That "|" in $\mathrm{P}(X|G)$ can be read as "given," and the whole expression can be read as "the probability of X given G."

Now the probability that X occurs given that G has already happened is the same thing as the probability that *both* X and G occurred, divided by the probability that G happened.

$$\mathrm{P}(X|G) = \frac{\mathrm{P}(X,G)}{\mathrm{P}(G)} \quad . \tag{1}$$

Of course if we can calculate the probability that X occurs given that G has already happened, we can also calculate the probability that G happened given that X occured:

$$\mathrm{P}(G|X) = \frac{\mathrm{P}(X,G)}{\mathrm{P}(X)} \quad . \tag{2}$$

(We can write $\mathrm{P}(X,G)$ in the equation (2) because it's just the probability of both X and G occurring. It doesn't make any difference what order they appear in.) If we multiply both sides of (1) by $\mathrm{P}(G)$ and both sides of (2) by $\mathrm{P}(X)$ we find

$$\begin{aligned} \mathrm{P}(X|G)\mathrm{P}(G) &= \mathrm{P}(X,G) \\ \mathrm{P}(G|X)\mathrm{P}(X) &= \mathrm{P}(X,G) \quad . \end{aligned}$$

So

$$\begin{aligned} \mathrm{P}(G|X)\mathrm{P}(X) &= \mathrm{P}(X|G)\mathrm{P}(G) \\ \mathrm{P}(G|X) &= \frac{\mathrm{P}(X|G)\mathrm{P}(G)}{\mathrm{P}(X)} \quad . \end{aligned} \tag{3}$$

Equation (3) is known as Bayes' theorem in honor of the Reverend Thomas Bayes, an 18th century English cleric who was the first to derive it.

The law of total probability

Bayes' Theorem is often used in calculating complicated probabilities by breaking the problem down into simple pieces. Suppose, for example, we wanted to calculate the probability that a father transmits the M allele to his daughter in a population. Fathers in the population may have one of three possible genotypes: MM, MN, or NN.

$$\begin{aligned} \mathrm{P}(X = M|G = MM) &= 1 \\ \mathrm{P}(X = M|G = MN) &= \frac{1}{2} \\ \mathrm{P}(X = M|G = NN) &= 0 \quad . \end{aligned}$$

To simplify the notation a little we'll write $\mathrm{P}(X)$ for $\mathrm{P}(X = MM)$, $\mathrm{P}(X|G_0)$ for $\mathrm{P}(X = M|G = MM)$, $\mathrm{P}(X|G_1)$ for $\mathrm{P}(X = M|G = MN)$, $\mathrm{P}(X|G_2)$ for $\mathrm{P}(X = M|G = NN)$, $\mathrm{P}(G_0)$ for $\mathrm{P}(G = MM)$, $\mathrm{P}(G_1)$ for $\mathrm{P}(G = MN)$, and $\mathrm{P}(G_2)$ for $\mathrm{P}(G = NN)$. The law of total probability tells us that

$$\mathrm{P}(X) = \mathrm{P}(X|G_0)\mathrm{P}(G_0) + \mathrm{P}(X|G_!)\mathrm{P}(G_1) + \mathrm{P}(X|G_2)\mathrm{P}(G_2)$$

In general,

$$\mathrm{P}(X) = \sum_i \mathrm{P}(X|G_i)\mathrm{P}(G_i) \quad ,$$

provided that the events represented by the G_i are mutually exclusive and exhaustive, i.e., that only one of them can happen and that one of them must happen.

Applying Bayes theorem

Let's return to our original problem, deciding whether the father of the young girl is homozygous for the M allele or heterozygous. Suppose instead of finding the value of G that maximimizes $\mathrm{P}(X|G)$ we want to calculate the probability that the father had the MM genotype. The probability we want is $\mathrm{P}(G = MM|X = M) = \mathrm{P}(G_0|X)$ and using Bayes' theorem, we can get it rather easily.

$$\mathrm{P}(G_0|X) = \frac{\mathrm{P}(X|G_0)\mathrm{P}(G_0)}{\mathrm{P}(X)} \quad .$$

Since we know the frequency of M in the population, 1%, and we can assume that genotypes are in Hardy-Weinberg proportions, we can calculate the

probability that a randomly chosen man in this population is homozgous for M, namely 0.01%. That's $\mathrm{P}(G)$. Now we can also calculate $\mathrm{P}(X)$ using the law of total probability

$$\begin{aligned}
\mathrm{P}(X) &= \mathrm{P}(X|G_0)\mathrm{P}(G_0) + \mathrm{P}(X|G_!)\mathrm{P}(G_1) + \mathrm{P}(X|G_2)\mathrm{P}(G_2) \\
&= (1 \times 0.0001) + \left(\frac{1}{2} \times 0.0198\right) + (0 \times 0.9801) \\
&= 0.01 \quad .
\end{aligned}$$

It shouldn't be too surprising that $\mathrm{P}(X)$, the probability that a randomly chosen father transmits the M allele to his daughter is equal to the frequency of that allele in the population, 1%. Putting this altogether we find that the probability that the father was homozygous for the M alllele is

$$\begin{aligned}
\mathrm{P}(G_0|X) &= \frac{1 \times 0.0001}{0.01} \\
&= 0.01 \quad .
\end{aligned}$$

We can also calculate the probability that the father was heterozygous or homozygous N.

$$\begin{aligned}
\mathrm{P}(G_0|X) &= \frac{1/2 \times 0.0198}{0.01} \\
&= 0.99 \\
\mathrm{P}(G_0|X) &= \frac{0 \times 0.9801}{0.01} \\
&= 0.00 \quad .
\end{aligned}$$

Comparing $\mathrm{P}(G_0|X)$ and $\mathrm{P}(G_1|X)$ we find that it is 99 times more likely that the father was heterozygous than that he was homozygous for M (and, not surprisingly, that there's no chance he was homozygous for N).