A note about expectations

For those not used to thinking about estimators and their properties, the idea of a biased MLE can be hard. Hopefully this note will help.

The general idea

When we have data, we make an estimate (such as the MLE). For a given data set this is just a number. But when we talk about the properties of the estimator, such its expectation, or its variance, we are talking about what values we might have got had we repeated this same sampling/experiment a large number of times. On average, what might we have got: this is biasedness/unbiasedness. How variable would the values be: this is summarized by the standard error, which is the square root of the variance of the estimator. These things are functions of the unknown parameter q. However, we just have one estimate of q from our data: to have a number for how variable things might be, we plug our value of q into the formula, to given ourselves an estimate of how variable our estimate might be.

The Binomial example: estimating accuracy

Back to counting alleles at a locus with 2 codominant alleles A and B. We wish to estimate the frequency of the A allele, q, from a sample of n individuals. Assuming HWE, this is just a sample of $2n$ independent alleles, and the count of A alleles is $B(2n, q)$ – this designates the binomial distribution with index $2n$ and parameter q. The proportion of A alleles is the count divided by $2n$, and the books tell us this has expectation $(2nq)/(2n) = q$ and variance $q(1-q)/2n$. What does this mean?

Well, I took a sample of $n = 20$ individuals from a population with $q = 0.2$, but in a real sample I do not actually know $q = 0.2$. (Actually, I got the computer to generate the sample.) I got 9 A alleles in my sample, so I estimate $\widehat{q} = 9/40 = 0.225$. How accurate do I think this estimate is? I have a formula $q(1-q)/40$ for the variance. If I plug in $q = 0.225$, I get 0.00436.

The standard error is then the the square root of this is 0.066, and generally we think we should be within 2 standard errors of the true value (or 3 if we want to be really sure). So here we can say that the real q is likely between $2.25 - 2 \times 0.066 = 0.093$ and $2.25 + 2 \times 0.066 = 0.357$. This is not saying much!! – but note we have only a small sample of 20 individuals. If instead I sampled 2000 individuals, and ended up with the same estimate of $q = 0.225$, then the estimated variance is 0.0000436 and standard error is 0.0066, and I could say that the real q is likely between $2.25 - 2 \times 0.0066 = 0.212$ and $2.25 + 2 \times 0.0066 = 0.238$. Note now we miss the true value 0.2! Another way of saying this is that if the real q is 0.2 and I sample 2000 individuals, I should not get as high a proportion as 0.225 of A alleles.

Unbiasedness and variation

So now the properties of my estimator relate to what happens if the experiment is repeated many times. I got the computer to sample 20 individuals (40 alleles), 10,000 times. The largest number of A alleles in this set of 10,000 samples was 18 and the smallest was 0. Those two unlucky samples would estimate $\widehat{q} = 18/40 = 0.45$ and $\widehat{q} = 0$ respectively. Clearly they did not get a good answer, but these are just the extremes out of 10,000 repetitions. Tne mean value of the number of A alleles over the 10,000 samples was 8.066, or the average estimate of q is 0.2016 – very close to 0.2. This is what unbiasedness means, on average we get the right answer. Note that even with 10,000 repetitions we do not get exactly 0.2, but that is enough to get very close.

The variance of my 10,000 counts of A alleles was 6.3396, which corresponds to a variance of estimates of q of 0.00396, or a standard error of 0.0629. Note that this is a bit smaller than the estimate of 0.066 I got from my single sample, but we were in the right ballpark. Note also the true theoretical value is $\sqrt{0.2 \times 0.8/40} = 0.0632$, but I can only say that if I know the true $q = 0.2$. However, my 10,000 repetitions got me pretty close. So, again, the estimate of variance I got from my single sample is telling me how variable the estimates are likely to be if the sampling is repeated, and hence how close I think the real q should be to my estimate.

An example of a biased MLE

Now let's take the same example of a sample of 20 individuals, but now suppose that the A allele is recessive.

Again suppose the true q is 0.2. Suppose I see just 1 AA individual in my sample, then the MLE of q is $\widehat{q} = \sqrt{1/20} = 0.223$ – not bad – I was lucky! My estimate of q^2 is 1/20 or 0.05.

Now the computer repeats this sampling of 20 individuals 10,000 times. For each individual in each of the samples the chance of being AA is $0.2^2 = 0.04$. In fact, in my 10,000 samples, the maximum number of AA is 6 (corresponding to an estimate $\widehat{q} = \sqrt{6/20} = 0.5477$ – way too large!! On the other hand, there were many of the 10,000 samples in which there were no AA individuals among the 20. In fact the average number of AA individuals is 0.7985, or the average proportion of AA individuals is $0.7985/20 = 0.0399$, or almost exactly the true value 0.04. That is, the proportion of AA individuals is an unbiased estimator of q^2.

However, our estimates are the square roots of these proportions. If we take the average of these over the 10,000 samples, it comes out at 0.1458 – quite different from the true value of $q = 0.2$. Even though the variance is quite large (0.01865), this certainly suggests this estimator is biased (as in fact it is).

So I took 500,000 samples, each of 20 individuals – good thing our computers get faster and more memory every year! The average count of AA individuals was still about 0.7985, and the average and the variance of the estimates (the square roots of the proportion of AA individuals) are almost exactly as before (0.1457, and 0.01868). So it looks like we have an accurate estimate of the true expectation here. That is, the expectation of the estimator is 0.1457, and not $q = 0.2$. The estimator is biased.