Notes on Schraiber et al., 2012 - *Genomic Tests on Variation in Inbreeding Among Individuals and Among Chromosomes*

Vince Buffalo

August 21, 2013

1 Levene's 1949 Model: The Probability of Heterozygotes with No Inbreeding

First, we consider equation (1) in Schraiber et al. (2012), the probability of seeing some number heterozygotes in a sample with no inbreeding derived from (Levene, 1949) (but I went back to Haldane (1954) and saw the same derivation there, earlier). The setup is that we have n diploid individuals, giving us n possible "bins" to put $2n$ balls into. If we only consider polymorphic sites (assuming biallelic), there are two alleles: ancestral and derived. For our $2n$ alleles in our sample, call i the number of derived alleles. i is bound between 1 and $2n-1$ (that is, $i \in \{1, 2, \ldots, 2n-1\}$) since if $i = 0$, we have no derived alleles and our site is not polymorphic, and if $i = 2n$, our site is entirely derived alleles, and again, our site is not polymorphic.

There are three additional constraints. First, if we define h as the number of heterozygous individuals (that is, individuals with one of the i derived alleles and one of the $2n-i$ ancestral alleles), then if h is odd, i must be odd, and second, if h is even, then i must be even. If this is unclear, consider the case where h is even. If an additional derived allele is added, it either goes into a heterozygous individual, and makes it homozygous (so $h' = h - 1$, making h odd) or it makes a homozygous individual heterozygous (so $h' = h + 1$, making h odd). Finally, the last constraint is that $h \leq \min(i, 2n - i)$, which can be understood as the most heterozygotes you can have is the minimum number of derived or ancestral alleles. This is easy intuitively too: in the extreme case that if *every* minor allele (i or $2n - i$) was in a heterozygous individual, then the most heterozygous genotypes possible is the number of minor alleles.

1.1 Combinatorics

Schraiber uses a classic colored ball and bin combinatorial approach equation borrowed on Levene (1949) and Haldane (1954) to look at the probability of heterozygotes given our seen number of derived alleles, i, and our sample size, n. Combinatorial approaches are very powerful, so I will go into some depth in this derivation.

When deriving probabilities using combinatorial approaches, our goal is to enumerate all outcomes in the sample space to determine the denominator, and then derive an expression for the number of outcomes for a particular event as the numerator. So to derive the probability of h heterozygotes given we see i derived alleles in a sample of n, we first want to determine the denominator by asking how many total ways are there are to arrange $2n - i$ red balls (ancestral alleles) and i white balls (derived alleles).

One way of thinking about this is to take all arrangements of $2n$ balls, treating each ball as distinguishable (there are $(2n)!$ of these). Since we don't care about the arrangement of indistinguishable white balls, we divide the number of arrangements of i white balls ($i!$). We also don't care about the number of arrangements of $2n - i$ indistinguishable red balls, so we divide out $(2n - i)!$ red balls. This leaves us with:

$$\frac{(2n)!}{i!(2n-i)!} \tag{1}$$

arrangements of i white balls and $2n - i$ red balls. Now, let n_2, n_1, and n_0 be the number of bins with 2, 1, and 0 white balls. Immediately, notice that $n_1 = h$, the number of heterozygotes, and $n_0 + n_1 + n_2 = n$.

Since i is the number of white balls (derived alleles), the number of n_0 bins (ancestral homozygous, bins with only red balls) is $n_0 = (2n - i - h)/2$, or intuitively, the number of ancestral (red) balls $(2n - i)$, minus the number in heterozygotes (h), leaving the number of ancestral balls in homozygous bins. Since we want a count of the number of *bins*, and not *balls*, and these are by definition in homozygous state, we divide by 2.

Next, for n_2, the number of homozygous derived (bins with two white balls) we use the same logic. There are i white balls, but we lose h of these two homozygous bins. Thus, we are left with $(i - h)/2$ bins that are homozygous derived.

Finally, n_2, n_1, and n_0 give us the counts for each of these events, as a function of i, h, and n. With these components, we can now get the number of distinguishable ways to assign *pairs* of balls to n bins. This is:

$$\frac{2^{n_1} n!}{n_2! n_1! n_0!} \tag{2}$$

which can be intuitively understood as there are $n!$ distinct sequences of the pairs, but since we don't care about the $n_2!$ ways of permuting the homozygous derived bins, we divide by those (and the same for $n_1!$ and $n_0!$). Note too that since each of the n_1 heterozygote bins has two arrangements (i.e. (w, r) and (r, w), and we don't care about the difference), we correct for this with 2^{n_1}. This leaves Equation 3, which is simply a rearranged Equation 2 over Equation 1.

$$\mathrm{P}(h \mid i, n) = \frac{i!(2n-1)!}{(2n)!} \frac{2^h n!}{((i-h)/2)! h! ((2n-i-h)/2)!} \tag{3}$$

2 Single Inbreed Coefficient

Next, Schraiber looks at a single inbreed coefficient for the sample, F. He uses F as the probability of being identical by descent and $1-F$ as the complement of this (I think Malécot was the first to see F this way). A probabilistic interpretation of F requires $0 \leq F \leq 1$ (no excess of heterozygotes).

Schraiber assumes that the j individuals at a polymorphic site that are i.b.d is binomially distributed with $p = F$. Thus:

$$\mathrm{P}(j \mid n', F) = \binom{n'}{j} F^j (1-F)^{n'-j} \tag{4}$$

Which is a standard binomial, with one minor caveat. Notice that we use n' here instead of n. This is to handle a constraint: if i (the number of derived alleles) is odd, then there has to be at least one heterozygous site. We define $n' = n$ if i is even, $n = n - 1$ if i is odd.

Now, let k be the number of individuals that are homozygous for the derived allele because they are i.b.d (given we know j). k has a hypergeometric distribution:

$$\mathrm{P}(k \mid j, i', n') = \left[\binom{i'/2}{k} (n' - i'/2j - k) \,/ \binom{n'}{j} \right] \tag{5}$$

This is a hypergeometric distribution, giving the probability of a k draws of success (autozygosity) with $i'/2$ "successes" (in our case, the number of derived alleles) in the population of size n', given j, our number of i.b.d. alleles in the *sample*.

Given j (the number of i.b.d. alleles and k (the number of derived i.b.d. allels), the number of individuals heterozygous at a site has the distribution given by Equation 3 with $n = n - j$ and $i = i - 2k$ (the paper says $j = j - 2k$ but there's no j in equation (1), so I believe this is a typo). Summing over j and k gives us the density:

$$\mathrm{P}(j \mid i, n, F) = \sum_{j=0}^{n'} \sum_{k=0}^{j} P(h, i-2k, n-j) P(j \mid F, n') P(k \mid j, i', n') \tag{6}$$

This equation can then be used as a likelihood for each derived allele i, by "replacing n by n_i, the observed number of sites at which there are i derived alleles, and h by h_i, the number of sites that are heterozygous." Why would estimating the probability of i.b.d. per each i ($\hat{F}_i$) in the sample be useful? If i is low (there are few derived alleles in the sample), it is probably young (think coalescent theory). Deviation from HWF in these may indicate selection.

Or, a single F for all i. To do this, we assume each i are independent and multiple them to get a MLE for F. A likelihood ratio test can be used to test whether $\hat{F}_i$ are equal.

3 Individual Specific Inbreeding Coefficient

Rather than just estimate F_i, one could estimate F_j, the inbreeding coefficient per individual $j \in \{1, \ldots, n\}$ (note this a different j than before!). Then, we can find the set $(F_1, \ldots, F_n)$

given i and $(h_1, \ldots, h_n)$ where $h_j = (h_{j1}, h_{j2}, \ldots, h_{jL})$ in which $h_{jl} = 1$ if individual j is heterozygous at site l and 0 otherwise (note: don't we need i_l?).

The exact likelihood is hard, so we compute the *composite likliehood.* Composite likelihoods are also known as pseudo-likelihoods because they don't fully specify the *real* likelihood (probabilistic model). Borrowing an example similar to one in Pawitan 2013, assume you have a linear spatial distribution of something and you observe some array of data, $y_1, y_2, \ldots, y_n$. The joint probability we reach for would be: $p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2 \mid y_1) \ldots p(y_n \mid y_{n-1}, \ldots)$. However, this imposes a right to left model of our spatial distribution, and seems awkward. It would be more natural to think of first-order neighbor model $p(y_k \mid y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_n) = p(y_k \mid y_{k-1}, y_{k+1})$. But, the product of all the conditional probabilities is not the same as the likelihood, but instead a composite likelihood. Pawitan gives a nice example for the simple case where observe (y_1, y_2):

$$p(y_1 \mid y_2)p(y_2 \mid y_1)$$

This is not a likeluhood; the true likelihood is:

$$p(y_1)p(y_2 \mid y_1)$$

The former case is the composite liklihood, as is similar to the one used in Schraiber et al.:

$$\mathrm{P}(h_1, h_2, \ldots, h_n \mid F_1, F_2, \ldots, F_n, i) = \prod_{j=1}^{n} \prod_{l=1}^{L} \mathrm{P}(h_{jl} = 1 \mid F_j, i_l)^{h_{jl}} \left(1 - \mathrm{P}(h_{jl} = 1 \mid F_j, i_l)\right)^{1-h_{jl}} \tag{7}$$

where L is the number of polymorphic sites (note I added an l subscript, as this is at a particular site and I think this is a typo in the original). Also:

$$\mathrm{P}(h_j = 1 \mid F_j, i) = (1 - F_j)\frac{i(2n - i)}{n(2n - 1)} \tag{8}$$

Note that for some individual j's F_j, this probability of being heterozygous has the same $1 - F$ factor as the frequency of a heterozygote under Generalized Hardy-Weinberg $((1 - F)2p(1 - p))$. So we might naturally ask, what is $(i(2n - i))/(n(2n - 1))$? Well, the frequency of derived alleles is $i/(2n - 1)$ and the frequency of ancestral is $(2n - i)/2n$. I believe the frequency of i has the denominator $2n - 1$ because the condition mentioned earlier.

References

JBS Haldane. An exact test for randomness of mating. *Journal of Genetics*, 52(3):631–635, 1954.

Howard Levene. On a matching problem arising in genetics. *The Annals of Mathematical Statistics*, 20(1):91–94, 1949.

Joshua G Schraiber, Stephannie Shih, and Montgomery Slatkin. Genomic tests of variation in inbreeding among individuals and among chromosomes. *Genetics*, 192(4):1477–1482, November 2012.