1.3.1 A SAMPLE OF GENES

- Consider a single genetic locus, with two codominant alleles A and B.
- Suppose each independent gene has allelic type A with probability q. We say q is the (population) allele frequency of allele A.
- For a random sample of n genes from the population, the number of A alleles is T ~ Bin(n,q).
- That is Pr(T=t) is proportional to q^t (1-q)^{n-t}.
- The obvious estimator of q is T/n.
- This estimator is unbiased since E(T/n) = nq/n = q.
- Its variance is q(1-q)/n which in fact is the smallest possible variance for any unbiased estimator.

1.3.2 Likelihood estimation of q

- The log-likelihood is λ(q)= t log (q) + (n-t) log(1-q).
- So differentiating the log-likelihood
- λ'(q) = (t/q) - (n-t)/(1-q) = n/(q(1-q))((t/n) – q)
- So the maximum likelihood estimator (MLE) is t/n.
- Differentiating again, we find the second derivative:
- λ''(q) = - (t/q^2) - (n-t)/(1-q)^2 and
- \- E(λ''(q)) = n/q + n/(1-q) = n/(q(1-q))
- This is the Fisher information, and the (large-sample) variance of the MLE is -1/E(λ''(q)).
- Here, q(1-q)/n is the variance for any sample size.
- For large n, MLEs are approx unbiased, and have approx the smallest possible variance.

1.3.3 A SAMPLE OF INDIVIDUALS

- Suppose we sample n individuals, and that n1 have genotype AA, n2 have genotype AB and n3 have genotype BB. n1+n2+n3 =n.
- Then we have (2n1 +n2) genes of allelic type A , in a sample of 2n genes.
- We can estimate q by (2n1 +n2)/2n, but properties of the estimator depend on the model for genotype frequencies:
- The log-likelihood is
  n1 log(P(AA)) + n2 log(P(AB)) + n3 log(P(BB)).

1.3.4 Four examples

- (i) The two genes in an individual must be of the same allelic type (n2=0): complete dependence. The estimator is n1/n and in effect we have a sample of n genes.
- (ii) Hardy-Weinberg equilibrium (HWE); independence of the allelic types of the two genes within an individual. So P(AA) = q^2, P(AB)= 2q(1-q) and P(BB) = (1-q)^2.
- (iii) A mixture of (i) and (ii): see 2.2.4.
- (iv) A mixture of subpopulations in HWE: see 1.3.5.

1.3.5 POPULATION STRUCTURE

- Suppose populations i proportions αi , each in HWE, with qij the freq of allele Aj in population i.
- The overall allele frequencies are weighted average of subpopulation allele frequencies.
- The overall genotype freqs are weighted average of subpopulation HWE frequencies.
- We can show that overall there is excess of each homozygote relative to overall HWE. This excess is known as the Wahlund variance.
- We can show that in total there are fewer heterozygotes than under HWE.
- Details of equations are on the next page.

Genotype frequencies under population structure:

First $\Pr(A_j) = q_{\cdot j} = \sum_i \alpha_i q_{ij}$, and so

$$\begin{aligned}\Pr(A_jA_j) - (\Pr(A_j))^2 &= \sum_i \alpha_i q_{ij}^2 - q_{\cdot j}^2 \\ &= \sum_i \alpha_i (q_{ij} - q_{\cdot j})^2 \geq 0 \\ \Pr(A_jA_k) - 2\Pr(A_j)\Pr(A_k) &= 2(\sum_i \alpha_i q_{ij} q_{ik} - q_{\cdot j} q_{\cdot k}) \\ &= 2\sum_i \alpha_i (q_{ij} - q_{\cdot j})(q_{ik} - q_{\cdot k})\end{aligned}$$

For two alleles, let $q_{i1} = q_i$, $q_{i2} = 1 - q_i$, $q_{\cdot} = q$.
If $\sigma_f^2 = \sum_i \alpha_i (q_i - q)^2$, then the three genotype freqs are $q^2 + \sigma_f^2$, $2q(1-q) - 2\sigma_f^2$ and $(1-q)^2 + \sigma_f^2$.

1.4.1 ESTIMATION: case of HWE

- Log-likelihood is $\lambda(q) = \log L(q)$
  $= n1 \log(q^2) + n2 \log(2q(1-q)) + n3\log((1-q)^2)$
  $= (2\,n1 + n2) \log(q) + (n2 + 2n3) \log(1-q)$
- The MLE of q is (2 n1 + n2)/2n.
- If T = 2 n1+n2, T ~ Bin(2n,q). --- back to binomial sampling, with a sample size 2n genes.
- Hence, var(T/2n) = q(1-q)/2n.
- Note: One generation of random mating establishes HWE, since, by definition, the two genes in an individual are copies of independently sampled parental genes.

1.4.2 Case of a recessive allele

- t = n1 of type AA, and n-t not of type AA.
- Assuming HWE, P(AA) = q^2, so log-likelihood is $\lambda(q) = t \log(q^2) + (n-t) \log(1-q^2)$
- Differentiating $\lambda'(q) = 2t/q - 2(n-t)\, q/(1-q^2)^2$
  . $= (2/q(1-q^2))\,(t - n q^2)$
- So the MLE of q is $\sqrt{t/n}$.
- Why should this be expected?
- Now T ~ Bin(n, q^2), but how can we find the variance of this MLE?

1.4.2 ctd: Using Fisher Information

- $\lambda''(q) = -2t/q^2 - 2(n-t)/(1-q^2)$
  . $- 4(n-t)\, q^2 / (1-q^2)^2$.
- $E(-\lambda''(q)) = 2n + 2n + 4 q^2 n/(1-q^2)$
  . $= 4n/(1-q^2)$
- Thus, the variance of the MLE of q is approx. (1-q^2)/4n.
- Note this is larger than q(1-q)/2n.
- Note (i) We have to make assumptions (HWE),
  . (ii) the variance of the estimator is larger.
  . (iii) Using the Fisher information we can measure the information lost.

1.4.3 Data on relatives

- We consider just mother-baby pairs and assume HWE.
- See next page for the conditional and joint probabilities.
- $l(q) = n00 \log(q^3) + n01 \log(q^2 (1-q)) +$
  $n10 \log(q^2 (1-q)) + n11 \log(q(1-q)) + n12 \log(q(1-q)^2)$
  $+ n21 \log(q(1-q)^2) + n22 \log((1-q)^3)$
  $= (3\,n00 + 2\,(n01 + n10) + n11 + n12 + n21) \log q +$
  $(3\,n22 + 2\,(n21+n12) + n11 + n10 + n01) \log(1-q)$
  $= mA \log q + mB \log(1-q)$.
- The MLE of q is mA/(mA +mB),
  where (mA +mB) = 3n - n11 and
  mA = (3 n00 + 2 (n01 +n10) + n11 + n12 + n21).

Parent and child probabilities

par	prob	ch AA	ch AB	ch BB		
AA	q^2	q	(1-q)	0		
AB	2q(1-q)	q/2	1/2	(1-q)/2		
BB	(1-q)^2	0	q	(1-q)		
	ch AA	ch AB	ch BB	Data counts		
AA	q^3	q^2(1-q)	0	n00	n01	0
AB	q^2(1-q)	q(1-q)	q(1-q)^2	n10	n11	n12
BB	0	q(1-q)^2	(1-q)^3	0	n21	n22

1.4.4 Alternatives to the MLE

- The MLE is ``best'', but there are simpler estimators that are not bad.
- One is to use only founders (here the moms): estimate q by (2 nAA + nAB)/2n where nAA and nAB are the numbers of AA and AB moms., (nAA = n00+n01).
- Or, use everyone, disregarding relationship: estimate q by (2 mAA + mAB)/4n, where mAA and mAB are is total numbers of AA and AB individuals. (mAA = 2 n00 + n01 + n10).
- These are both unbiased estimators, but asymptotically the MLE has smaller variance.