BCB 568 Homework Assignment 2

Schuyler Smith

February 8, 2018

6. The Hardy-Weinberg Principle is a model for genetics in a population that uses allele frequencies to explain genotype ratios. It has 5 essential asssumptions required for it to be applicable to a diploid sexually reproducing population:

1. Random mating
2. Infinite population size
3. No Mutation
4. No migration
5. No selection

This model allows us to know that at a loci the bi-allelic frequencies (p and q) sum to 1.

$$1 = p + q$$

Arguably more importantly, it demonstrates that the genotype ratios for a population under these assumptions are binomially distributed.

$$1 = p^2 + 2pq + q^2$$

When these hold true and the observed genotypes for the population occur in the modeled ratios it is said that the population is in Hardy-Weinberg Equilibrium (HWE).
If we assign the refernce allele for a populations $\psi = p$ we can say that the genotype G for i individuals is randomly distributed to $Bin(m, \psi)$

$$P(G_i = g|\psi) = \binom{m}{g} \psi^g (1 - \psi)^{m-g}$$

7. If we take all of the assumptions required for HWE to hold true, and all the assumptions for the model of genotype probabilites from Part I, then our likelihood function would be

$$\begin{aligned} L(\psi|\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n) &= P(\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n|\psi)P(\psi) \\ &= \prod_{i=1}^{n} P(\boldsymbol{d}_i|\psi)P(\psi) \\ &= \prod_{i=1}^{n} \sum_{g=0}^{m} P(\boldsymbol{d}_i|G=g,\psi)P(G=g|\psi) \end{aligned}$$

8. **a.** The log likelihood of the function from 7 give us

$$\begin{aligned} \log L(\psi|\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n) &= \log \prod_{i=1}^{n} \sum_{g=0}^{m} P(\boldsymbol{d}_i|G=g,\psi)P(G=g|\psi) \\ &= \sum_{i=1}^{n} \log \sum_{g=0}^{m} P(\boldsymbol{d}_i|G=g,\psi)P(G=g|\psi) \end{aligned}$$

The included figure (plot_log_by_psi), shows the log-likelihood of using the provided next-gen read data.

b. optim() is an R function that can be used to perform Brent optimization. The default mode of optim is to minimize a function, to maximize it we set the fnscale = -1.
The R-code to perform this and any subsequent operations is included in a seperate file.
Our outpus from optim gives us $\hat{\psi} = 0.749$ as our maximum likelihood estimate (MLE) for the reference allele frequency for the population.

c. If we use ψ_0 to be the true ratio of reference alleles in the population then our MLE will be consistent with $\hat{\psi}$ as our sample size increases. The nature of MLEs allow us to say

$$\hat{\psi} \sim N\Big(\psi_0, \frac{1}{nI(\psi_0)}\Big)$$

with

$$I(\psi_0) = E\Big(\frac{\partial^2 \log L}{\partial \psi^2}\Big)$$

So this would mean a strategy for estimating the variance of $\hat{\psi}$ is to take enough samples for asymptotic normality to hold, and then calculating the Fisher information $I(\psi_0)$ for $\hat{\psi}$.

9. $S = \{d_1, d_2, \ldots, d_{10}\}$ is the 10 individuals in the original data set and we want to use bootstrapping by resampling from S 10 times with replacement. Using $\boldsymbol{d}_i^*$ to represent the individual i of the resample we can use Brent optimization again to find $\hat{\psi}*$

$$\hat{\psi}* = \arg\min \log L(\psi|\boldsymbol{d}_1^*, \boldsymbol{d}_2^*, \ldots, \boldsymbol{d}_{10}^*).$$

10. We resample our ngs data 10,000 times and create a histogram from the observations (bootstrap_estimates). A 90% confidence interval can be found by sorting the data of the 10,000 boostrap resamples and using the tails as a cutoff, making us 90% confident that the true value of ψ lies between the sorted $\hat{\psi}*$ values of the 501^{th} to 9500^{th}.
This gives us a $CI = [0.60, 0.90]$

11. The 90% confidence interval obtained from the likelihood ratio test statistic is $CI = [0.57, 0.88]$. This is slightly different than the bootstrapping method which has given a higher estimate for the interval, I'm more inclined to believe the bootstrapping method centers closer to the actual value of ψ, but that's purely intuition, I have absolutely no quantitative argument to back that up. Truly, both CIs are so wide that neither is really informative.