Stat 536 Homework 5

Due: 10/3/08

In this question, you will derive an estimator for effective population size in populations where only a small fraction of the total population is allowed to reproduce. This situation is typical in domesticated animals, where reproduction is limited to a few individuals, while the rest serve another purpose. We will assume that the breeding pool is drawn at random from the larger population, with equal numbers of males and females (which is probably not so typical for domesticated animals). Assuming the breeding pool is randomly mating according to HWE assumptions, the effective population size is equal to the size of the breeding pool N_b, but let's suppose we cannot observe the size and want to estimate it. The key observation that leads to an estimator is that there will be an *excess* of heterozygotes in the offspring, and this excess heterozygosity can be easily measured. In what follows, assume no selfing and distinguish the sexes so that offspring must have one mother and one father.

1. **Recall the homework 2 question that claimed when allele frequencies differ in the sexes by amount δ, then heterozygote frequency in the offspring is**

$$H = 2pq + \frac{\delta^2}{2}$$

where p and $q = 1 - p$ are the allele frequencies in the offspring. Previously, δ was viewed as fixed. In this context δ is a random variable that depends on the sampling process that created the breeding population, so to use this equation, we'll substitute

$$H = 2pq + \frac{E(\delta^2)}{2}$$

Suppose the total population has allele frequencies p_0 and q_0. By considering the variance in allele proportions in small, random subpopulations (breeding populations) of size N_b (males and females both number $\frac{N_b}{2}$), show that the above equation can be rewritten as

$$H = 2pq + \frac{p_0 q_0}{N_b} \tag{1}$$

Solution:

The breeding process randomly *and independently* selects $\frac{N_b}{2}$ females and $\frac{N_b}{2}$ males from the large population. Assume the large population is currently at HWE so that sampling $\frac{N_b}{2}$ individuals is the same thing as sampling N_b alleles. Then, the allele frequencies in the male p_m and female p_f breeding populations are binomial probabilities and will differ because of sampling variability. Their expectation and variance are for $i \in \{f, m\}$

$$E(p_i) = p_0 \qquad\qquad \mathrm{Var}(p_i) = \frac{p_0 q_0}{N_b}$$

The difference $\delta = p_f - p_m$ has mean 0 and variance

$$\mathrm{Var}(\delta) = \mathrm{Var}(p_f) + \mathrm{Var}(p_m) = \frac{2p_0q_0}{N_b}$$

by independence. Also, $\mathrm{Var}(\delta) = E(\delta^2)$, so plugging the above back in for $E(\delta^2)$ gives the result.

2. From equations in the lecture one can show that the ratio of "heterozygosity" across two subsequent generations is

$$\frac{h_{t+1}}{h_t} = \left(1 - \frac{1}{2N_e}\right)$$

where N_e is the effective population size for any non-Wright-Fisher population. Relate h_t and h_{t+1} to expected heterozygosity under HWE and use this observation to eliminate p_0 and q_0 from the right-hand-side of eq. (1), leaving

$$H = 2pq\frac{N_b + \sqrt{N_b^2 + 1}}{N_b - 1 + \sqrt{N_b^2 + 1}}$$

[Note: $N_e \neq N_b$. What is N_e for your population that breeds via a breeding population of size N_b?]

Solution:

$h_t = 1 - f_t$ is the probability of non-IBD individuals, i.e. the proportion of heterozygotes *plus* the proportion of non-IBD homozygotes in the population. We cannot distinguish IBD and non-IBD homozygotes in our sample. However, heterozygotes are lost at the same rate as non-IBD homozygotes, so

$$\frac{h_{t+1}}{h_t} = \frac{P_{12}(t+1)}{P_{12}(t)}$$

We now equate $P_{12}(\cdot)$ with expected heterozygosity under HWE ($2pq$). The allele frequency in the first generation is p_0 and in the second generation p, so

$$\frac{h_{t+1}}{h_t} = \frac{P_{12}(t+1)}{P_{12}(t)} = \frac{2pq}{2p_0q_0} = \frac{pq}{p_0q_0} \tag{2}$$

Disregard all but the breeding population for the moment. The lecture notes state that for a non-selfing, balanced-sex population of size N_b

$$\frac{h_{t+1}}{h_t} = \left(1 - \frac{1}{2N_e}\right) = \frac{1 - \frac{1}{N_b} + \sqrt{1 + \frac{1}{N_b^2}}}{2}$$

Thus,

$$\begin{aligned} H &= 2pq + \frac{pq}{N_b}\frac{p_0 q_0}{pq} \\ &= 2pq + \frac{pq}{N_b}\left(\frac{2}{1-\frac{1}{N_b}+\sqrt{1+\frac{1}{N_b^2}}}\right) \\ &= 2pq + \frac{2pq}{N_b - 1 + \sqrt{N_b^2+1}} \\ &= 2pq\frac{N_b + \sqrt{N_b^2+1}}{N_b - 1 + \sqrt{N_b^2+1}} \end{aligned}$$

3. H **is the observed heterozygosity.** $2pq$ **is the HWE predicted heterozygosity. Define**

$$D = \frac{H - 2pq}{2pq}$$

as a measure of heterozygote excess and show that

$$N_b = \frac{1}{2D} + \frac{1}{2(D+1)}$$

Solution:

From here, it is just algebra. Let

$$E = \frac{H}{H - 2pq}$$

and note

$$E = \frac{1}{D} + 1 \qquad\qquad 1 - \frac{1}{E} = \frac{1}{D+1}$$

$$\begin{aligned} H &= 2pq\frac{N_b + \sqrt{N_b^2+1}}{N_b - 1 + \sqrt{N_b^2+1}} \\ H(N_b + \sqrt{N_b^2+1}) - H &= 2pq(N_b + \sqrt{N_b^2+1}) \\ \frac{H}{H-2pq} - N_b &= \sqrt{N_b^2+1} \\ E - N_b &= \sqrt{N_b^2+1} \\ E^2 - 2N_bE + N_b^2 &= N_b^2 + 1 \\ E - 2N_b &= 1/E \\ N_b &= \frac{E - 1/E}{2} \end{aligned}$$

4. **Find an estimate of N_b and confidence intervals by bootstrap for the following data. Please check your bootstrap datasets carefully and deal sensibly with any problems encountered. Dealing "sensibly" means you solve the problem somehow and you at least make mention of it in your solutions.**

AA	Aa	aa
135	**502**	**363**

Solution:

5. **Critique this estimator. Why might it not be a very good estimator?**

Solution:

We made many approximations and assumptions in the derivations above.

- In particular, we replaced δ^2 in the equation for offspring heterozygote frequency with its expectation. The actual difference in allele frequencies between the sexes may vary from near zero to very large; it will especially vary as N_b becomes small.
- We have assumed that the heterozygote frequencies in both generations satisfy HWE. We assume HWE for the preceding generation to answer part 1 and we assume it for both generations to derive the equation in part 2. The fact that HWE, in fact, does not apply because of genetic drift is absolutely true and integral to the derivation.

Furthermore, when applied to the dataset, it seems that our estimator had high variance (especially when considering the negative values it yielded). In part, this is a result of the form of the estimator. D, which can be quite small by chance, even with small N_b, appears in the denominator and can make the estimator explode. In addition, it is not hard to see that random fluctuation can give an invalid estimate $N_b < 0$ even when the true N_b is well above 0, another sign of high variance. All in all, there is a lot of noise in this estimator.

Including data from multiple alleles could reduce the noise, but HWD could result for many different reasons, and any one of them could be violated and misinterpretted by this method as caused by a small breeding population.