Contents

3.5 Estimating N_e 1
3.5.1 Introduction 1
3.5.2 Temporal-Based Estimation of N_e 2

3.5 Estimating Effective Population Size

3.5.1 Introduction

Review: N_e

Recall that N_e for a real population Z is the size of a hypothetical Wright-Fisher population Y that would give the same value of *some genetic property* as in the real population Z.

The *genetic property* that is matched can vary so there are different effective population sizes: inbreeding, variance, eigenvalue, mutation, or coalescent, but the most popular are:

- N_{eI}: inbreeding effective population size matches rate of decrease in heterozygosity h_{t+1}/h_t
- N_{eV}: variance effective population size matches variance in allele frequency change $p_{t+1} - p_t$

Fortunately, $N_{eI} = N_{eV}$ if the population is at equilibrium, i.e. when all reproductive, demographic, and environmental effects, including the effective population size, are constant. Additionally, they are asymptotically equivalent $N_{eI}, N_{eV} \approx N_e$, where N_e is the harmonic mean of (varying) effective population size when the population is not at equilibrium. Henceforth, we will merely refer to N_e.

Biological Theory

There are a number of reasons why being able to estimate N_e is valuable.

- N_e interacts with other forces (mutation, selection, migration, recombination) to determine the genetic variation in a population, so knowing N_e tells us something about the evolutionary mechanisms that *acted* on a population to produce the present-day variability.
- N_e can be also be used to predict:
  - loss and distribution of neutral variation
  - fixation probability of new alleles (new mutations, slightly deleterious or advantageous)
  - fitness and survival of small populations
- To facilitate design of breeding schemes, where N_e is presumably controlled by the breeder.

General Findings/Conventions

Estimates of N_e have been made based on data collected from many different natural populations. There are two general observations:

- $N_e < N$, where N is the census size of the population; in fact one study found that *all* 192 publications they examined found this inequality to be true.
- $N_e \approx 0.1 \times N_e$

Conservation biology, which is concerned with conserving populations in healthy states, at least according to Paetkau1998, follows these rules of thumb:

- Make every attempt to keep $N_e \geq 50$.
- Target $N_e \in [500, 5000]$ for the long-run health of the population.

Methods for Estimating N_e

- **Demography-Based.** Last lecture we discussed variations on the Wright-Fisher model and derived N_e for a few special cases. The resulting formula provide equations where N_e is given in terms of demographic parameters, for example

$$N_e = \frac{4N_f N_m}{N_f + N_m} + \frac{1}{2}$$

  One can use these equations to directly estimate N_e from known demographic parameters (in this example, N_m and N_f). Unfortunately, it can be difficult to estimate demographic parameters, so these methods are little used. See [**?**] for a review.

- **Genetic Methods.** The remaining methods are based on collecting genetic information from populations.
  - **Phylogenetic-Based.** Infer historical relationships between multiply sampled individuals, and use the shape of the resulting tree to infer N_e. We can't discuss these methods until we discuss how to model and infer historical relationships among individuals. These methods provide *long-term* estimates of N_e since they are based on long historical relationships. The remaining methods provide *short-term* estimates of N_e.
  - **Disequilibrium-Based.** Random fluctuations in populations introduce small associations that register as non-zero disequilibria when equilibria are expected. These methods take a single sample from a population and are based on genotype and allele sample proportions.
    - **Heterozygouos excess.** Theory suggests that more heterozygotes than one would expect under HWE arise in finite populations. The resulting excess $P_{12} - 2p_1p_2$ is estimated from data and related to N_e through theoretical equations. There is open research to utilize the variation in genotype frequencies from HWE predictions (not just heterozygote excess) to get better estimators.
    - **Linkage disequilibrium.** Measurements of linkage disequilibrium D_{AB} and related quantities can be also related to N_e and inverted for estimation.
  - **Temporal-Based.** Finite populations are expected to experience changes in allele frequencies when comparing across generations. When data can be collected from multiple generations, variation in the allele frequencies can be estimated and related to N_e using theory. We will explore this method in greater detail to demonstrate how genetic data can be converted into $\hat{N}_e$. Similar procedures are used for the other genetic methods, see Schwartz1998 and Wang2005 for reviews.
  - **Equilibrium in Subdivided Populations.** These methods we will discuss later in the course after we have discussed population subdivision and how genetic drift balances against other evolutionary forces.

3.5.2 Temporal-Based Estimation of N_e

Background

Recall

$$\text{Var}(p_{t+1}) = p_1(1-p_1)\left[1 - \left(1 - \frac{1}{2N_e}\right)^t\right] \tag{1}$$

I have slightly edited this equation to match what follows, but note that the initial allele frequency p_1 is the frequency of the allele in the adults of generation 1. By definition of variance and because $E(p_t) = p_1$ for all

$t > 1$, we also have $\text{Var}(p_{t+1}) = E\left[(p_{t+1} - p_1)^2\right]$. Now, this quantity $(p_{t+1} - p_1)^2$ is something we could estimate if we sample from generation 1 and $t+1$ (e.g. plug in MLEs $\hat{p}_{t+1}$ and $\hat{p}_1$). With estimates in hand, we could potentially invert eq (1) to come up with an estimator for N_e.

There is a complication though. Suppose we observe K alleles and estimate $(p_{k,t+1} - p_{k1})^2$ for each of them. Unfortunately, each one will have a different expectation [eq. (1)] that depends on p_{k1}, so it is not obvious how to combine multiple alleles. Instead, note

$$f = \frac{\text{Var}(p_{t+1})}{p_1(1-p_1)} = 1 - \left(1 - \frac{1}{2N_e}\right)^t$$

is the same for all alleles. Thus if instead we estimate

$$f_k = \frac{(p_{k,t+1} - p_{k1})^2}{p_{k1}(1-p_{k1})} \tag{2}$$

then we can take the sample mean across alleles and get an even better (lower variance) estimate of f

$$f = \bar{f}_k = \frac{1}{K}\sum_{k=1}^{K} \frac{(p_{k,t+1} - p_{k1})^2}{p_{k1}(1-p_{k1})}$$

Multiple Estimators

And in fact, this is exactly the first proposed estimator. Define

$$\hat{F}_a = \frac{1}{K}\sum_{k=1}^{K} \frac{\left(\tilde{p}_{k,t+1} - \tilde{p}_{k1}\right)^2}{\tilde{p}_{k1}\left(1 - \tilde{p}_{k1}\right)}$$

Unfortunately, the above estimator can become infinite if $\tilde{p}_{k1} = 0$, but $\tilde{p}_{k,t+1} > 0$, which can happen by chance due to sampling. A second proposed estimate is

$$\hat{F}_c = \frac{1}{K}\sum_{k=1}^{K} \frac{\left(\tilde{p}_{k,t+1} - \tilde{p}_{k1}\right)^2}{(\tilde{p}_{k1} + \tilde{p}_{k,t+1})/2 - \tilde{p}_{k1}\tilde{p}_{k,t+1}}$$

This latter estimator is justified by noting that the denominator in eq. (2) is $p_1 - p_1^2$ (dropping the dependence on allele k). If there is no mutation, selection, migration, etc., then $E(p_t) = p_1$, so

$$E\left(\frac{\tilde{p}_1 + \tilde{p}_{t+1}}{2}\right) = p_1 \qquad \text{and} \qquad E(\tilde{p}_{t+1}\tilde{p}_1) \approx p_1^2$$

the latter by neglecting possible dependence between $\tilde{p}_{t+1}$ and $\tilde{p}_1$ (which should be weak, as we'll see). The end result is the denominator in $\hat{F}_c$ is still estimating what we desired:

$$E\left[(\tilde{p}_{k1} + \tilde{p}_{k,t+1})/2 - \tilde{p}_{k1}\tilde{p}_{k,t+1}\right] = p_1 - p_1^2$$

What Comes Next

Now, we will determine $E(\hat{F})$ by assuming a theoretical model. This expectation will be a function of quantities we know and some we don't, notably the population parameter N_e. Thus, Method of Moments can be used to produce an estimate $\hat{N}_e$ by inverting the resulting equation.

$$E(\hat{F}_c) = g(N_e, \ldots) \quad \text{becomes usable via Method of Moments as:} \quad \hat{F}_c = g(N_e, \ldots)$$

for some invertible function $g(\cdot)$ so that $\hat{N}_e = g^{-1}(\hat{F}_c, \ldots)$.

Model

Assume the following population and properties

- diploid
- random mating population of size N
- discrete generations
- no selection, migration, mutation

Then, sample S_1 individuals at time 1 and S_{t+1} individuals at generation $t+1$ to produce sample proportions $\tilde{p}_1$ and $\tilde{p}_{t+1}$ of an allele of interest.

Expectation of $\hat{F}_c$

We need to formally compute the expectation of our statistic.

$$E\left(\hat{F}_c\right) = \frac{1}{K}\sum_{k=1}^{K} E\left(\frac{\left(\tilde{p}_{k,t+1} - \tilde{p}_{k1}\right)^2}{(\tilde{p}_{k1} + \tilde{p}_{k,t+1})/2 - \tilde{p}_{k1}\tilde{p}_{k,t+1}}\right)$$

so it is a sum of expectations of the following form

$$E\left(\frac{(x-y)^2}{(x+y)/2 - xy}\right) \approx \frac{E[(x-y)^2]}{E[(x+y)/2 - xy]} \tag{3}$$

where x is the sample proportion at the first sampling time and y is the sample proportion at the next sampling time, t generations later. Of course, it is not technically legal to conclude the expectation of a ratio is the ratio of expectations, but it is approximately true, and we will use this approximation to continue.

$E(\hat{F}_c)$

Let p_0 be the true (and unknown) allele frequency in the generation immediately preceding the first sampled generation 1. Clearly, $E(x) = E(y) = p_0$ and $E(x-y) = 0$, so $E[(x-y)^2] = \text{Var}(x-y)$ and the numerator of our expectation is

$$\text{Var}(x-y) = \text{Var}(x) + \text{Var}(y) - 2\text{Cov}(x,y)$$

the denominator is

$$\begin{aligned} E[(x+y)/2 - xy] &= p_0 - [E(xy) - p_0^2] - p_0^2 \\ &= p_0(1-p_0) - \text{Cov}(x,y) \end{aligned}$$

which involves again the covariance term.

We will now find $\text{Var}(x)$, $\text{Var}(y)$, and $\text{Cov}(x,y)$ one at a time.

Var(x)

First, sampling individuals from generation 1 in a population with the assumptions we have made is like sampling from the gamete pool of generation 0. Therefore,

$$\text{Var}(x) = \frac{p_0(1-p_0)}{2S_1}$$

Var(y)

As for Var(y), we remember

$$\text{Var}(p_t) \quad = \quad p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\right]$$

and argue, as for Var(x), that given the allele frequency p_t of the preceding generation

$$\text{Var}(y \mid p_t) \quad = \quad \frac{p_t(1-p_t)}{2S_{t+1}}$$

Then

$$\begin{aligned}
\text{Var}(y) &= E\left[\text{Var}(y \mid p_t)\right] + \text{Var}\left[E(y \mid p_t)\right] \\
&= E\left[\frac{p_t(1-p_t)}{2S_{t+1}}\right] + \text{Var}(p_t) \\
&= \frac{p_0}{2S_{t+1}} - \frac{1}{2S_{t+1}}E(p_t^2) + \text{Var}(p_t) \\
&= \frac{p_0}{2S_{t+1}} - \frac{1}{2S_{t+1}}\left[\text{Var}(p_t) + p_0^2\right] + \text{Var}(p_t) \\
&= \frac{p_0(1-p_0)}{2S_{t+1}} + \left(1-\frac{1}{2S_{t+1}}\right)p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\right] \\
&= p_0(1-p_0)\left[1-\left(1-\frac{1}{2N_e}\right)^t\left(1-\frac{1}{2S_{t+1}}\right)\right]
\end{aligned}$$

Cov(x, y)

[This slide is not fully justified, but there is some confusion in the literature over the form of Cov(x, y).]

If the sample from generation 1 is taken without replacement before reproduction, then Cov$(x, y) = 0$ because x and y derive from two independent samples from generation 0 gamete pool. [I also think this makes the assumption that N_e is not affected by the sampling process, but these are small details when N is rather large compared to the sample size S_0.]

Unfortunately, if the sample is taken with replacement or after reproduction from a finite population, then there could be overlap and covariation between the sample taken from generation 1 and the "sample" of size N_e taken to make the next generation. To see how much covariation there is, write

$$\begin{aligned}
x &= p_1' + x - p_1' \\
y &= p_1' + y - p_1'
\end{aligned}$$

where p_1' is the allele frequency in the total generation 1 population of size N. Then, [no details shown]

$$\text{Cov}(x, y \mid p_1') \quad = \quad \text{Var}(p_1') = \frac{p_0(1-p_0)}{2N}$$

Var$(x - y)$

So, overall, the variance is

$$\begin{aligned}
\text{Var}(x-y) &= p_0(1-p_0)\left[\tfrac{1}{2S_1} + 1 - \left(1-\tfrac{1}{2N_e}\right)^t\left(1-\tfrac{1}{2S_{t+1}}\right) - \tfrac{1}{N}\right] && \text{sample with replacement} \\
\text{Var}(x-y) &= p_0(1-p_0)\left[\tfrac{1}{2S_1} + 1 - \left(1-\tfrac{1}{2N_e}\right)^t\left(1-\tfrac{1}{2S_{t+1}}\right)\right] && \text{sample without replacement}
\end{aligned}$$

$E(\hat{F}_c)$

And putting the numerator and denominator back together

$$E(\hat{F}_c) = \frac{\text{Var}(x-y)}{p_0(1-p_0) - \text{Cov}(x,y)} = \frac{\text{Var}(x-y)}{p_0(1-p_0)(1-1/(2N))}$$

where the $1/(2N)$ term in the denominator is there only when sampling with replacement and is otherwise 0. However, $1/(2N)$ is usually very small for reasonably sized populations and it is ignored yielding

$$E(\hat{F}_c) \approx \frac{\text{Var}(x-y)}{p_0(1-p_0) - \text{Cov}(x,y)} = \frac{\text{Var}(x-y)}{p_0(1-p_0)}$$

Then, approximately

$$\begin{aligned} E(\hat{F}_c \mid \text{sample with replacement}) &= \frac{1}{2S_1} + 1 - \left(1 - \frac{1}{2N_e}\right)^t \left(1 - \frac{1}{2S_{t+1}}\right) - \frac{1}{N} \\ E(\hat{F}_c \mid \text{sample without replacement}) &= \frac{1}{2S_1} + 1 - \left(1 - \frac{1}{2N_e}\right)^t \left(1 - \frac{1}{2S_{t+1}}\right) \end{aligned}$$

When $t/(2N_e)$ is small (i.e. t small or N_e large), then the expectations can be simplifed farther

$$\begin{aligned} E(\hat{F}_c \mid \text{sample with replacement}) &\approx \frac{1}{2S_1} + \frac{1}{2S_{t+1}} + \frac{t}{2N_e} - \frac{1}{N} \\ E(\hat{F}_c \mid \text{sample without replacement}) &\approx \frac{1}{2S_1} + \frac{1}{2S_{t+1}} + \frac{t}{2N_e} \end{aligned}$$

$\hat{N}_e$

Inverting these equations, finally leads to estimates

$$\begin{aligned} \hat{N}_e &= \frac{t}{2[\hat{F}_c - 1/(2S_1) - 1/(2S_{t+1}) + 1/N]} && \text{sample with replacement} \\ \hat{N}_e &= \frac{t}{2[\hat{F}_c - 1/(2S_1) - 1/(2S_{t+1})]} && \text{sample without replacement} \end{aligned}$$

If the census population size N is large, then there is little difference between these estimates. If there is some concern about dropping the $1/N$ term, but no census estimate is available, then one could use the roughly repeatable observation that $N_e \approx 0.1N$ and re-arrange the equation to solve for N_e again.

Choosing t

In order to use the estimators, some value for the number of generations t separating sample times must be available. One usually estimates t not from data, but from additional knowledge about the species in question. For example, if the two sample timepoints 1 and $t+1$ are S years apart, then $t \approx \frac{S}{G}$, where G is the average generation length of the animal in question. For example, for humans $G \approx 25$ might be appropriate (although it could vary with country). If there is some uncertainty about t, you would identify a plausible range (a kind of confidence interval) and generate a range of possible N_e's, e.g.

$$\hat{N}_e \in \left[\frac{t_L}{2[\hat{F}_c - 1/(2S_1) - 1/(2S_{t+1})]}, \frac{t_U}{2[\hat{F}_c - 1/(2S_1) - 1/(2S_{t+1})]} \right]$$

where t_L and t_U are the identified plausible limits for t.

Multiple Loci

When there are multiple loci, then a weighted average across loci is used

$$\hat{F}_c = \sum_j \frac{K_j \hat{F}_{cj}}{\sum_i K_i}$$

where K_j is the number of alleles observed at locus j.

Confidence Intervals

It has been noticed that the distribution of $\hat{F}_c$ appears rather like a chi-squared distribution. Formally,

$$\frac{n\hat{F}}{E(\hat{F})} \sim \chi^2_n$$

where $n = \sum_j (K_j - 1)$ degrees of freedom (total number of alleles removing the lost degree of freedom because allele frequencies must sum to one at each locus). This approximation seems good as long as there are no alleles with very low frequency.

Because the chi-squared distribution is quite skewed, it is not good to build confidence intervals by computing a variance and assuming an approximate normal (symmetric) distribution. In addition, $\hat{F}$ can be ∞ or negative for some formulations, which can cause problems in computing variances. Much better confidence intervals are given by

$$\left[\frac{n\hat{F}}{\chi^2_{1-\alpha/2}}, \frac{n\hat{F}}{\chi^2_{\alpha/2}}\right]$$

where $\chi^2_\cdot$ are quantiles of a chi-squared distribution with n degrees of freedom. [This is based on the observation that $\hat{F}$ is an estimator of variance and $\frac{n\hat{F}}{E(\hat{F})}$ is distributed like $\frac{ns^2}{\sigma^2}$ (for a sample of size $n+1$), which is the classic route at confidence intervals for population variance σ^2.]

Assumption Violations

- **Constant** N_e: Not a problem, but $\hat{N}_e$ approximates the harmonic mean of the true (varying) N_e.
- **Discrete Generations**: Overlap OK if the population is stable; if not the bias increases.
- **No Selection**: Constant selection has minor effect if t/N_e is small.
- **No Mutation**: Time scale of data collection makes this assumption reasonable.
- **No Migration**: Population structure can lead to strong biases. We will talk about population structure next.

References

Caballero, A. (1994) Developments in the prediction of effective population size. *Heredity.* **73**: 657–79.

Krimbas, C. B. and S. Tsakas (1971) The genetics of *Dacus oleae* V. Changes of esterase polymorphism in a natural population following insecticide control – selection or drift. *Evolution.* **25**: 454–60.

Pollak, E. (1983) A new method for estimating the effective population size from allele frequency changes. *Genetics.* **104**: 531–48.

Schwartz, M. K., Tallmon, D. A., and G. Luikart (1998) Review of DNA-based census and effective populationn size estimators. *Animal Conserv.* **1**: 293–9.

Wang, J. (2005) Estimation of effective population sizes from data on genetic markers. *Philos. Trans. Royal Soc. B.* **360**: 1395–409.

Waples, R. S. (1998) A generalized approach for estimating effective population size from temporal changes inallele frequency. *Genetics.* **121**: 379–91.