Estimating the Effective Number of Breeders From Heterozygote Excess in Progeny

Gordon Luikart* and Jean-Marie Cornuet²

*Laboratoire de Biologie des Populations d’Alti­tude, Université Joseph Fourier, CNRS, F-38041 Grenoble Cedex 9, France and
²Laboratoire de Modélisation et de Biologie Evolutive, INRA/URLB, 34090 Montpellier Cedex, France

Manuscript received June 17, 1998
Accepted for publication November 20, 1998

ABSTRACT

The heterozygote-excess method is a recently published method for estimating the effective population size (Nₑₑ). It is based on the following principle: When the effective number of breeders (Nₑₑ) in a population is small, the allele frequencies will (by chance) differ between males and females, which causes an excess of heterozygotes in the progeny with respect to Hardy-Weinberg equilibrium expectations. We evaluate the accuracy and precision of the heterozygote-excess method using empirical and simulated data sets from polygamous, polygynous, and monogamous mating systems and by using realistic sample sizes of individuals (15–120) and loci (5–30) with varying levels of polymorphism. The method gave nearly unbiased estimates of Nₑₑ under all three mating systems. However, the confidence intervals on the point estimates of Nₑₑ were sufficiently small (and hence the heterozygote-excess method useful) only in polygamous and polygynous populations that were produced by <10 effective breeders, unless samples included > ~60 individuals and 20 multiallelic loci.

The effective population size (Nₑₑ) is an important parameter in evolutionary genetics and conservation biology because it influences the rate of inbreeding and loss of genetic variation. It also influences the efficiency of natural selection in maintaining beneficial alleles and purging deleterious ones. For example, when Nₑₑ is very small, genetic drift will often be too strong for natural selection to efficiently maintain or purge alleles. Unfortunately, Nₑₑ has proven very difficult to estimate in natural populations (Waples 1989). Thus, any method with potential for improving our ability to estimate Nₑₑ deserves thorough evaluation. Nₑₑ can be estimated using demographic or genetic data. However, demographic methods require information such as variance in reproductive success, which is difficult to obtain for many species. Further, demographic estimates of Nₑₑ are often higher than the true Nₑₑ because demographic methods seldom incorporate all of the factors (e.g., skewed sex ratios, change in population size, etc.) that can make Nₑₑ smaller than the population census size (Nₑₑ) (Frankham 1995).

The (current) effective population size can be estimated using genetic data and the four following statistical methods: the loss of heterozygosity method (e.g., Harris and Allendorf 1989); the temporal method (Krimbas and Tsaskas 1971; Waples 1989); the linkage disequilibrium method (Hill 1981; Waples 1991); and, most recently, the heterozygote-excess method (Pudovkin et al. 1996). Pudovkin et al. (1996) demonstrated that the heterozygote-excess method estimates the effective number of breeding parents (Nₑₑ) with no bias and fair precision when the sample size of progeny is infinite and when gametes combine completely at random, i.e., when all male gametes have an equal probability of combining with all female gametes, as in some polygamous, random-mating species (e.g., marine invertebrates such as shellfish).

Here, we extend the evaluation of Pudovkin et al. (1996) to include finite samples of individuals (n = 15–120), reasonable numbers of polymorphic loci (5–30) with a wide range of allele frequencies, monogamous mating systems where only pair-mating occurs, and polygynous mating systems where only a few males mate with many females (i.e., skewed sex ratios). Our goal is to delineate the conditions under which the heterozygote-excess method will be useful for estimating Nₑₑ in natural populations (or in captive populations such as fish hatcheries). In this article, we (i) explain the importance of using an unbiased estimator of the expected heterozygosity (Hₑₑ) for calculating Nₑₑ from finite samples of progeny, (ii) quantify the bias of Nₑₑ, (iii) determine the number of loci and individuals that must be sampled to achieve precise estimates of Nₑₑ, and (iv) test if monogamy generates an interfamily Wahlund effect (i.e., heterozygote deficiency) that counteracts the heterozygote excess generated by small Nₑₑ. To conduct these evaluations, we use data from simulations and natural populations. We focus on populations with a
small \( N_{eb} (4-100) \) because a heterozygote excess is generated only when \( N_{eb} \) is small.

**PRINCIPLE AND METHODS**

The principle of the heterozygote-excess method is as follows: When the number of breeders is small, the allele frequencies in males and females will be different due to binomial sampling error. This difference generates an excess of heterozygotes in the progeny relative to the proportion of heterozygotes expected under Hardy-Weinberg equilibrium (Robertson 1965; Rasmussen 1979). The proportion of heterozygotes expected in progeny produced by a small and equal number of females and males can be calculated as (Falconer 1989, p. 67)

\[
H' = 2pq + pq/n = 2pq(1 + 1/(2n)),
\]

where \( n \) is the number of haploid genomes in the mothers or fathers and \( p \) and \( q \) are the frequencies of alleles at a locus, in the population from which the parents were drawn. Because the excess of heterozygotes expected in progeny increases as the number of parents decreases, Pudovkin et al. (1996) suggested using the magnitude of heterozygote excess to estimate the effective number of breeding parents.

Pudovkin et al. (1996) called \( H' \) the proportion of heterozygotes expected to be observed \( (H_{obs}) \) in the progeny (given a limited number of parents), whereas the expected proportion of heterozygotes in the base population under Hardy-Weinberg proportions, \( 2pq \), was designated as \( H_{exp} \).

Now \( N_{eb} \) can be estimated as follows:

\[
\hat{N}_{eb} = H_{exp}/(2(H_{obs} - H_{exp})).
\]

The ratio \( H_{obs}/(H_{obs} - H_{exp}) \) is the reciprocal of Sander's (1970) index \( D \), for excess or deficiency of heterozygotes; thus, an estimate of \( N_{eb} \) is (Pudovkin et al. 1996)

\[
\hat{N}_{eb} = 1/(2D).
\]

The following more exact equation was derived by Pudovkin et al. (1996):

\[
\hat{N}_{eb} = 1/(2D) + 1/(2(D + 1)).
\]

The above expression is for two alleles. For a multiallelic locus, one should average \( D \) over all \( k \) alleles as

\[
D = (1/k)(\sum D_i),
\]

where \( D_i \) is the excess of heterozygotes for the \( ith \) allele,

\[
D_i = (H_{obs(i)}/H_{exp(i)}).
\]

\( H_{obs(i)} \) is the total proportion of heterozygotes having allele \( i \), and \( H_{exp(i)} \) is simply \( 2p_i(1-p_i) \cdot 2N/(2N - 1) \), where \( i \) is the \( ith \) allele.

We note that when the sample size of individuals is finite, \( H_{exp} \) must be estimated using the following unbiased estimator of \( 2pq \) (Nei 1987): \( 2N(2pq)/2N - 1 \), where \( N \) is the number of progeny sampled. If \( 2pq \) is used instead of the unbiased estimator, \( N_{eb} \) will give a biased estimate \( N_{eb} \) (especially when \( N \) is small). In nature, \( N_{eb} \) can range from only two to near infinity and \( N_{eb} \) can be negative (e.g., in the case of an overall deficit of heterozygotes). In our analyses, we considered \( N_{eb} \) values as infinite if \( N_{eb} \) was negative or >10000 (an arbitrary but reasonably large value). If \( N_{eb} \) is infinite, it simply means that the heterozygote-excess signal is obscured by the noise (i.e., sampling error) associated with small samples of loci or individuals.

The simulation model that we used to evaluate the heterozygote-excess method has three main steps. First, it assigns genotypes to the parental generation using random numbers and a predefined allele frequency distribution. We modeled loci with 2, 3, and 5 alleles and the following three allele frequency distributions: equal (e.g., \( H = 0.8 \) for 5 alleles), triangular (e.g., 0.33, 0.30, 0.20, 0.13, 0.07, and \( H = 0.74 \); see Pudovkin et al. 1996), and rare (e.g., 0.52, 0.2, 0.10, 0.07, 0.04, with \( H = 0.67 \)). Second, the simulation model randomly picks a male and female parent and one gamete from each. Under the polygamy model, the gametes from males and females are

| TABLE 1 |
| Harmonic mean \( \hat{N}_{eb} \) and distribution of 95% confidence limits for \( \hat{N}_{eb} \) from 500 simulations |

| \( N_{eb} \) | Loci | n | \( N_{eb} \) | L | C | H | \( N_{eb} \) | L | C | H | \( N_{eb} \) | L | C | H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 10 30 | 4.1 | 1 | 92 | 7 | 4.3 | 1 | 96 | 5 | 4.2 | 6 | 87 | 9 | 4.0 | 3 | 86 | 11 |
| 4 10 120 | 4.0 | 1 | 93 | 6 | 4.2 | 1 | 93 | 6 | 4.2 | 6 | 84 | 10 | 4.0 | 5 | 84 | 11 |
| 4 20 30 | 4.2 | 1 | 92 | 6 | 4.0 | 2 | 94 | 4 | 4.0 | 5 | 84 | 11 | 4.2 | 6 | 84 | 10 |
| 4 20 120 | 4.0 | 2 | 92 | 6 | 4.2 | 1 | 93 | 6 | 4.2 | 6 | 84 | 10 | 4.0 | 5 | 84 | 11 |
| 20 10 30 | 22.2 | 3 | 95 | 2 | 22.2 | 4 | 94 | 2 | 21.3 | 5 | 84 | 11 | 21.9 | 6 | 89 | 5 |
| 20 10 120 | 20.4 | 3 | 94 | 3 | 20.9 | 3 | 94 | 3 | 21.9 | 6 | 89 | 5 | 21.3 | 5 | 84 | 11 |
| 20 20 30 | 22.8 | 5 | 92 | 3 | 22.2 | 4 | 92 | 4 | 21.8 | 8 | 85 | 7 | 21.9 | 6 | 89 | 5 |
| 20 20 120 | 21.4 | 4 | 92 | 4 | 20.4 | 4 | 92 | 4 | 20.0 | 8 | 83 | 9 | 20.4 | 4 | 92 | 4 |
| 100 10 30 | 103.4 | 1 | 94 | 5 | 110.3 | 2 | 95 | 3 | 95.4 | 9 | 83 | 8 | 106.8 | 2 | 94 | 4 |
| 100 10 120 | 102.1 | 2 | 94 | 4 | 114.3 | 3 | 94 | 3 | 113.5 | 8 | 83 | 9 | 106.8 | 2 | 94 | 4 |
| 100 20 30 | 102.8 | 2 | 94 | 4 | 105.6 | 3 | 94 | 3 | 99.4 | 8 | 84 | 8 | 112.6 | 8 | 84 | 8 |

Tri-5 and Equal-5 are loci with five alleles and a triangular or equal allele frequency distribution (i.e., 0.33, 0.30, 0.20, 0.13, 0.07 or 0.2, 0.2, 0.2, 0.2, 0.2). Tri-2 are loci with two alleles and a triangular allele frequency distribution (0.67, 0.34). \( n \) is the number of individuals sampled. \( \hat{N}_{eb} \) is the harmonic mean of 500 simulation estimates of \( N_{eb} \), L, C, and H show the percentage of 500 confidence intervals (for 500 independent \( \hat{N}_{eb} \) estimates) that were too low, that contained the true \( N_{eb} \), or that were too high.
This is repeated until 15±120 offspring have been generated. Not contain the true randomly chosen and gametes are drawn to make an offspring. Approximately half of the confidence intervals that did with a second random female. Then one of the two pairs is when using loci with five alleles (Table 1). As expected, 

\[ N_e = 4(\text{Even-5}), (\text{Tri-5}), (\text{Tri-3}), (\text{Tri-2}), (\text{Monogamy}), (\text{Polygamy}), (\text{Tri-5}) \]

This suggests that there is little impact of an interfamily Wahlund effect on the accuracy of \( N_e \) estimates. When only approximately two to three large families exist (e.g., under monogamy with \( N_e \) equaling 4–6), sampling across families does not generate a large heterozygote deficiency (i.e., Wahlund effect). However, a Wahlund heterozygote deficiency is expected when many families exist (A. Pudovkin, personal communication). Such a deficiency would at least partially cancel the heterozygote excess caused by small \( N_e \), and thereby cause biased (high) estimates of \( N_e \). Although monogamy did not cause a large bias, it did substantially increase the variance among \( N_e \) estimates (see below and Figure 1).

**Precision:** To quantify the precision of the \( N_e \) estimates, we used the Student’s t-distribution to compute a 95% confidence interval for each \( N_e \) (as in Pudovkin et al. 1996). The confidence interval on \( N_e \) contained the true \( N_e \) in 92–96% of the simulation estimates of \( N_e \) when using loci with five alleles (Table 1). As expected, approximately half of the confidence intervals that did not contain the true \( N_e \) were too low (L) and half were too high (H). This suggests that the Student’s t-distribution is useful for computing confidence intervals, even though \( N_e \) is not exactly normally distributed. For loci with three alleles or for monogamous mating systems, the confidence intervals also contained the true \( N_e \)≈92–96% of the time. However, when using loci with only two alleles, the confidence intervals were generally too narrow and contained the true \( N_e \) in only
83–89% of the simulation estimates of $N_{eb}$ (Table 1). Thus, confidence intervals must be interpreted with caution or computed by alternative methods (e.g., bootstrap resampling) when using loci with only two alleles (e.g., many allozyme loci).

Under extreme polygyny (e.g., one male mating with 99 females), the confidence intervals were often too high. For example, when $N_{eb}$ was four, ~25% of the 500 simulated confidence intervals were slightly higher than the true $N_{eb}$, and none were lower than $N_{eb}$. Although the confidence intervals were often too high, they were also much narrower under polygyny than under monogamy or polygamy (Figure 1). This narrowness substantially increases the usefulness of the heterozygote-excess method under polygyny. Thus, under extreme polygyny, the heterozygote-excess method will be useful for detecting a small $N_{eb}$ but will be less useful for quantifying the exact size of $N_{eb}$.

To determine the minimum number of loci and individuals that must be sampled to achieve a high probability of obtaining narrow confidence intervals, we plotted the distribution of the (upper and lower) 95% confidence limits obtained from 500 simulation estimates of $N_{eb}$. When the true $N_{eb}$ is only 4, at least 10 loci (with five alleles) and 30 individuals must be sampled to achieve an 80% probability of obtaining an upper 95% confidence limit $< 20$ (Figure 2b). In other words, the statistical power is 0.80 when testing the null hypothesis that the true $N_{eb} = 20$ (and when the true $N_{eb}$ is actually only 4). The power will be slightly higher when using a one-tailed test and the null hypothesis that true $N_{eb} \geq 20$ (the alternative hypothesis is $N_{eb} < 20$).

These results show that the heterozygote-excess method is sufficiently powerful for detecting a small $N_{eb}$ when sampling reasonable numbers of individuals and loci with five alleles. Such results are important for conservation biology and the management of captive and natural populations. The precision of $N_{eb}$ is often increased more by analyzing a larger number of loci than by sampling more individuals. Doubling the number of loci from 10 to 20 generally reduces confidence intervals more than doubling the number of individuals sampled from 15 to 30 (compare the first two box plots in Figure 2b). However, the benefit from doubling the number of loci depends on the number and frequency of alleles (Figure 1).

When true $N_{eb}$ is 10, we must sample >20 polymorphic loci and 60 individuals to have an 80% probability of obtaining confidence intervals that are $< 50$ (and to have a 95% probability of obtaining confidence intervals $< 100$; Figure 2e). When the true $N_{eb}$ is 100, >80% of the confidence intervals include infinity, even when sampling 120 individuals and 20 loci with five alleles (data not shown). Clearly, when $N_{eb}$ is >10, very large samples of loci and individuals are required to achieve a high probability of obtaining reasonably small confidence intervals. Thus, the main limitation of the usefulness of the heterozygote-excess method is its poor precision, i.e., its wide confidence intervals. The confidence intervals are generally too wide for the method to be useful when using diallelic loci, loci with mostly rare alleles, or when studying strictly monogamous species (Figure 1).

When applied to data from natural populations, the
heterozygote-excess method often gave estimates of $N_{eb}$ equal to infinity. For example, $N_{eb}$ was infinity in 5 of 10 cohorts for which the total number of parents was known (or estimated) to be small (i.e., three to a few dozen). Further, only 2 of the 10 estimates gave 95% confidence intervals that did not include infinity as an upper limit (Table 2). This poor precision is not surprising in that only 5–9 polymorphic loci were analyzed, and only 11–25 progeny were sampled. Additional empirical evaluations are needed, but it is extremely difficult to find large data sets containing individuals produced from a known number of parents.

One potential limitation of the method is the requirement for random, representative sampling. For example, if a sample contains only one or few families (due to sampling error) then we could obtain a very low $N_{eb}$ estimate, even though many families actually exist and $N_{eb}$ is large. Another obvious limitation is that the method will work only in species with separate sexes. The method will work for haplo-diploid species (e.g., Hymenopterans), but will require the derivation of equations different from those presented here.

Four approaches may help circumvent the problem of poor precision. First, one can compute 80% confidence intervals (in addition to 95% confidence intervals). This will reduce the likelihood that the upper confidence limit will include infinity and be uninformative. Second, one could explore alternative methods for computing confidence intervals (e.g., nonparametric methods such as bootstrap resampling of loci). Third, one could combine estimates of $N_{eb}$ from several generations or cohorts by computing the harmonic mean of $N_{eb}$ over the multiple generations or cohorts. This can reduce the probability of obtaining infinity for $N_{eb}$ because, when computing a harmonic mean, the low estimates carry far more weight than high ones (e.g., infinity). Finally, one can potentially combine estimates of $N_{eb}$ obtained from several independent $N_{e}$ estimators by computing the harmonic mean of the $N_{e}$ estimates. Other promising $N_{e}$ estimators include those based on gametic disequilibrium (Hill 1981) and on temporal variance in allele frequencies (Krimbas and Tsakas 1971; Waples 1991). These two estimators also suffer from low precision (Waples 1989, 1991; Luikart 1997; Luikart et al. 1998). However, two or more of the estimators may be independent (Waples 1991; Pudovkin et al. 1996) and thus could potentially be used simultaneously to achieve a more precise estimate of $N_{e}$. More research is needed to evaluate the precision and accuracy that can be achieved by using several $N_{e}$ estimators simultaneously. Any improvement in our ability to estimate $N_{e}$ would be significant in light of both the difficulties in assessing $N_{e}$ and the importance of $N_{e}$ in population genetics and in conservation biology.

We thank I. Till-Bottraud and two anonymous reviewers for helpful comments on earlier versions of this manuscript, M. Schwartz for sharing unpublished simulation data, and especially P. Spruell, F. W. Allendorf, A. Estoup, and M. Brown for providing data sets. Support was provided by the French “Bureau Ressources Génétiques,” a postdoctoral fellowship (for G.L.) from National Science Foundation/
North Atlantic Treaty Organization, and the Laboratoire de Biologie des Populations d’Altitude.

LITERATURE CITED


Communicating editor: G. B. Golding