Introduction to Population Genetics

F. A. (Phil) Aravanopoulos

Laboratory of Forest Genetics & Breeding, School of Forestry & Natural Environment, Aristotle University of Thessaloniki, Greece

SCOPE OF POPULATION GENETICS

Population genetics is the study of how Mendel's laws and other genetic principles apply to the entire populations

Such a study is essential to a proper understanding of evolution, because, fundamentally, evolution is the result of progressive changes in the genetic composition of a population

Aims of population genetics

Population genetics seek to understand and predict the effects on populations of such genetic phenomena as:

Segregation

Recombination

Mutation

taking into account such ecological and evolotionary factors as

Population size

Mating pattern

Geographic distribution of individuals

Migration

Natural selection

Population Genetics

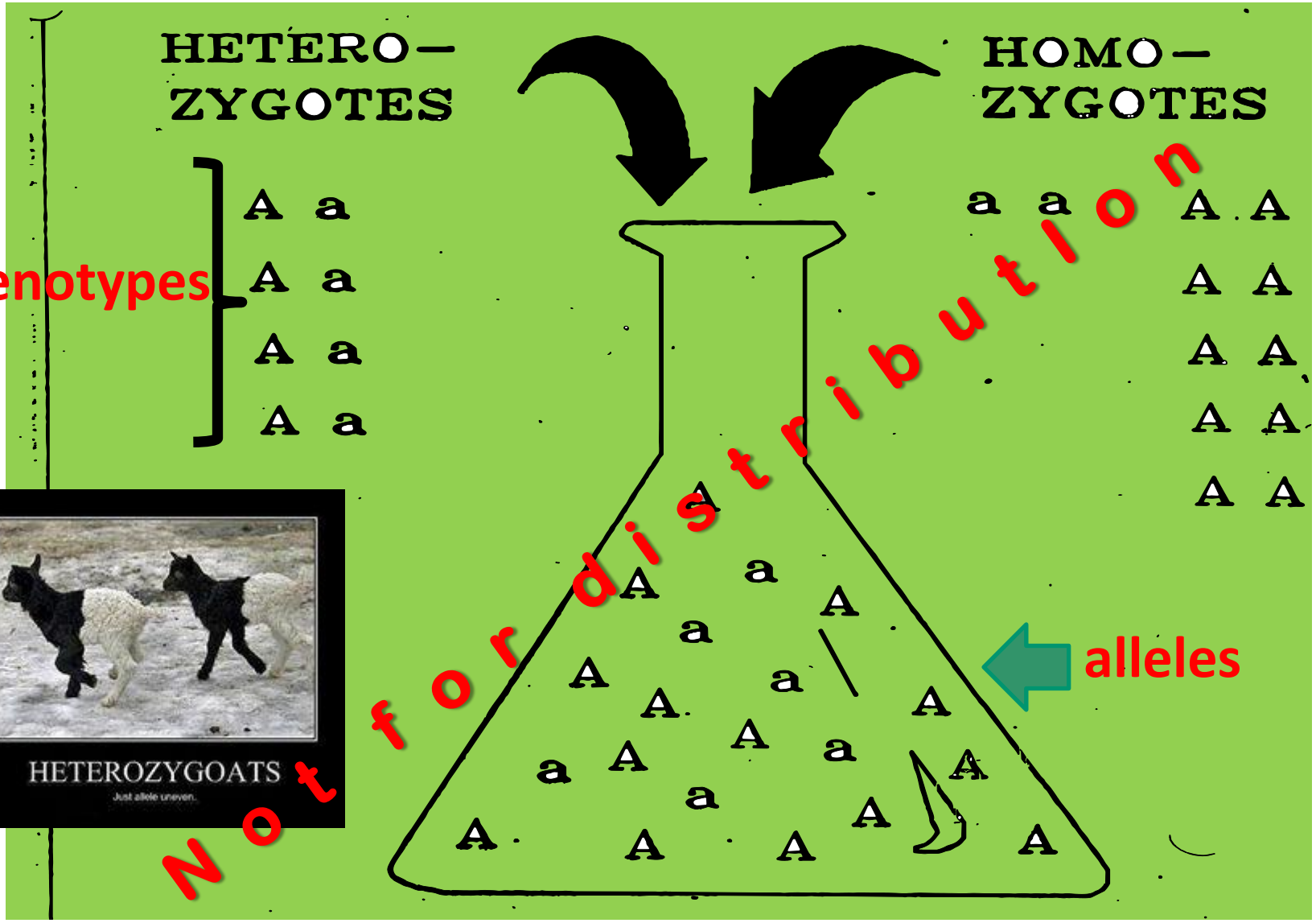
- Concerned with the distribution and generation to generation transmission of alleles within a population.
- Deals with genetic variation due to polymorphic loci, where more than one allele is present at those loci.
- Essence of Mendelian inheritance is directly applied to the population, controlling the distribution of genotypes within the population and the transmission of existing variation to subsequent generations.

Natural population

A group of individuals of a specific spatial reference that belong to the same species an can actually or potentially interbreed.

Genetic description of populations

- Necessary to have some convenient quantitative measure of genetic variation
- Basic measures:
  - Genotype frequency
  - Allele frequency
    - Allele = one of several alternative forms of a gene


HETERO–
ZYGOTES
HOMO–
ZYGOTES
genotypes
A a
A a
A a
A a
a a
A A
A A
A A
A A
A A
HETEROZYGOATS
Just allele uneven.
Not for distribution
alleles

GENE POOL

Panmictic Model

Assumptions made in developing the model for predicting genotype frequencies

1. Diploid organism
2. Sexual reproduction
3. Generations are non overlapping
4. Mating is random
5. Population size is infinite
6. No migration among populations
7. Mutation is extremely rare
8. All genotypes equally fit, no natural selection

These assumption constitute the Hardy Weinberg model

If all assumptions are met, allele and genotype frequencies would remain the same over time (over different reproductive cycles).

Measures of Genetic Diversity, Differentiation & Distance

Allele Frequencies

Allele frequencies are calculated using the observed genotypes in a sample:

$$\text{Freq(allele)} = \frac{(2 \times \text{Obs. Homozygotes}) + (\text{Obs. Heterozygotes})}{2 \times \text{Individuals Sampled}}$$

Can we derive genotype frequencies from allele frequencies?

- Two basic assumptions:
  - Mendel's law of independent segregation
  - Random mating (frequency of mating is given by frequency of genotypes involved)

Random mating of individuals = random union of gametes

Predicting genotype frequencies from knowledge of allele frequencies is quite straightforward, but there are few complications.

Genotype frequencies are determined in part by mating pattern. One of the simplest and most important mating pattern is random mating, in which mating takes place at random with respect to the gene under consideration.

With random mating, the chance that one individual mates with another having a prescribed genotype is equal to the frequency of that genotype in the population

Basic genetic diversity parameters

- Three basic :
- (1) percent polymorphic loci

$$P = N_P / r$$

(N_P:number of polymorphic loci, r: total number of loci

- (2) average number of alleles per locus (allelic richness
-

$$A = \Sigma m_i / r$$

(m_i: number of alleles of the *i^{th} locus*)

Basic genetic diversity parameters

- (3) average (expected) heterozygosity (under Hardy-Weinberg)

  (m:number of alleles f_i: frequency of the του i^{th} allele at a locus)

$$H_e = 1 - \sum_{i=1}^{m} (f_i)^2$$

POPULATION STRUCTURE

- We have focused on variation within a population, what about variation among populations?

- Populations in nature can be somewhat connected by migration, yet still be somewhat independent = **metapopulation**

Population Substructure

- Many species naturally subdivide themselves into herds, flocks, colonies, schools etc.
- Patchy environments can also cause subdivision
- Subdivision decreases heterozygosity and generates genetic differentiation via:
  - » Natural selection
  - » Genetic drift

Mean heterozygosities at population level

- Heterozygosity = mean percentage of heterozygous individuals per locus
- Assuming H-W, heterozygosity (H) = 2pq where p and q represent mean allele frequencies
- H_S = sum of all subpopulation heterozygosities divided by the total number of subpopulations

Not for distribution

Wright's Fixation Index

- Equals the reduction in heterozygosity expected with random mating at one level of population hierarchy relative to another more inclusive level.

$$F_{ST} = (H_T - H_S)/ H_T$$

POPULATION STRUCTURE

- We commonly measure population structure using fixation indices or **F-statistics.**
- We can measure among population variation using Fst

$$Fst = \sigma_q / [\bar{q}(1-\bar{q})]$$

variance in allele frequency

average allele frequency among populations

Interpreting F_{ST}

- Can range from 0 (no genetic differentiation) to 1 (fixation of alternative alleles).
- Wright's Guidelines:
  - 0 - 0.05, little differentiation
  - 0.05 – 0.15, moderate
  - 0.15 – 0.25, great
  - > 0.25, very great

G_{ST}

- G_{ST}: **multi-allelic analogue of F_{ST}** (Nei 1986,1987)

$$G_{ST} = D_{ST}/H_T = (H_T - H_S)/H_T$$

where D_{ST} is the average gene diversity between subpopulations

G_{ST}

- $D_{ST} = H_T - H_S$
- $G_{ST} = D_{ST} / H_T$
- $H_T = 1 - G_{ST}$
- G_{ST} shows the proportion of total genetic diversity that resides within populations

θ

- θ: **unbiased estimator of Fst** that corrects for error associated with incomplete sampling of a populations (Weir and Cockerham 1984)

$$\hat{\theta} = \frac{a}{a + b + c},$$

- a = between pop variance, b= between individuals within pops, c= between gametes within individuals

R_{ST}, G_{ST}, and θ

- R_{ST}: explicitly accounts for mutation rates at **microsatellite loci** (Slatkin 1995)

$$R_{ST} = (S_T - S_W)/S_T$$

- R_{ST} is the fraction of the total variance of allele size that is between populations
- S is the avg. sum of squares of difference in allele sizes

Not for distribution

AMOVA (Analysis of Molecular Variance)

- Method of estimating population differentiation directly from molecular data (e.g. RFLP, direct sequence data, or phylogenetic trees)
- The variance components are used to calculate phi-statistics which are analogous to Wright's F-statistics

$$\Phi_{ST} = (\sigma^2_a + \sigma^2_b)/\sigma^2_T$$

Genetic Distance (D)

- Quantitative measure of genetic divergence between two sequences, individuals, or taxa

- Relative estimate of the time that has passed since two populations existed as a single, panmictic population

- Units of *D* depend on the kind of molecular data collected (allozymes, nucleotide sequences, etc.)

Two Most Commonly used Distance Measures

- Nei's genetic distance (Nei, 1972)
- Cavalli-Sforza chord measure (Cavalli-Sforza and Edwards, 1967)

- Nei's assumes that differences arise due to mutation and genetic drift, C-S and RWC assume genetic drift only

Nei's Genetic Distance

- $D = -\ln I$

  where $I = \Sigma x_i y_i \,/\, (\Sigma x_i^2 \, \Sigma y_i^2)^{0.5}$

- For multiple loci, use the arithmetic means across all loci
- Interpreted as mean number of codon substitutions per locus

Assumptions for Nei's Distance

- All loci have same rate of neutral mutation
- Mutation-genetic drift equilibrium
- Stable effective population size
- IAM

Not fo[illegible] distribution

Cavalli-Sforza Chord Distance

- populations are conceptualized as existing as points in a **m**-dimensional Euclidean space which are specified by **m** allele frequencies (i.e. **m** equals the total number of alleles in both populations). The distance is the angle between these points:

$$d_{chord}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{P} x_{1j} x_{2j}}{\sqrt{\sum_{j=1}^{P} x^2_{1j} \sum_{j=1}^{P} x^2_{2j}}}\right)}$$

- xi and yi are the frequencies of the **i**th allele in populations **x** and **y**
- Assumes genetic drift only (no mutation)
- Geometric distance b/w points in multi-dimensional space

Testing Significance of Distance Measures

- *Bootstrap*: generation of many new data sets by resampling original data with replacement.
- For each bootstrap data set, obtain estimates of parameters of interest and their variances
- Generates confidences intervals of parameter estimates.

Relaxation of the Panmictic Model: Genetic Drit

GENETIC DRIFT

- Reduction in population size, especially substantial reductions, initiate the process of random genetic drift
- In this process, a restricted and variant sample of the genes present in the parental population survives into the next generation
- The random changes in allelic frequencies that occur due to sampling error, including the loss of alleles, are called ***random genetic drift***

- When a large populations reduced in size such as during a bottleneck, genetic drift becomes important because of the two following main effects:

  - ***Loss of alleles***
  - ***Erosion of heterozygosity, or genetic variance***

Loss of alleles

The expected loss depends on the distribution of allele frequencies

Allele frequency distributions range from being uniform or "even", with allele equally frequent, to highly skewed, with few frequent alleles and many rare ones

For loci with similar allelic richness (number of alleles), the loss of alleles resulting from random genetic drift is much less when distributions are even than is the loss from skewed distribution

Not for distribution

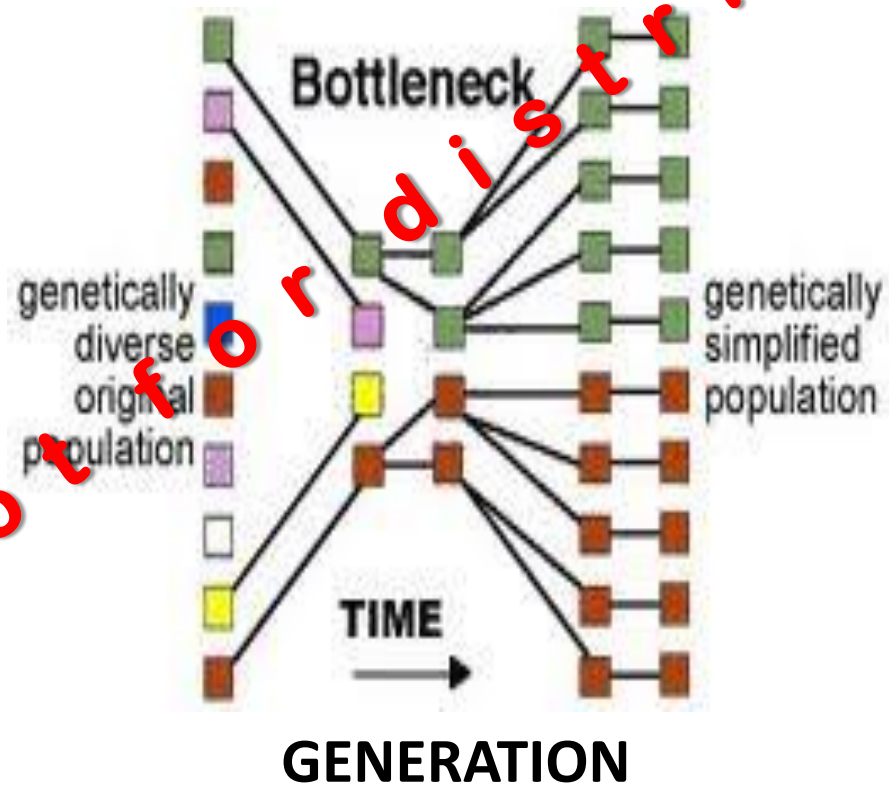
POPULATION SIZE

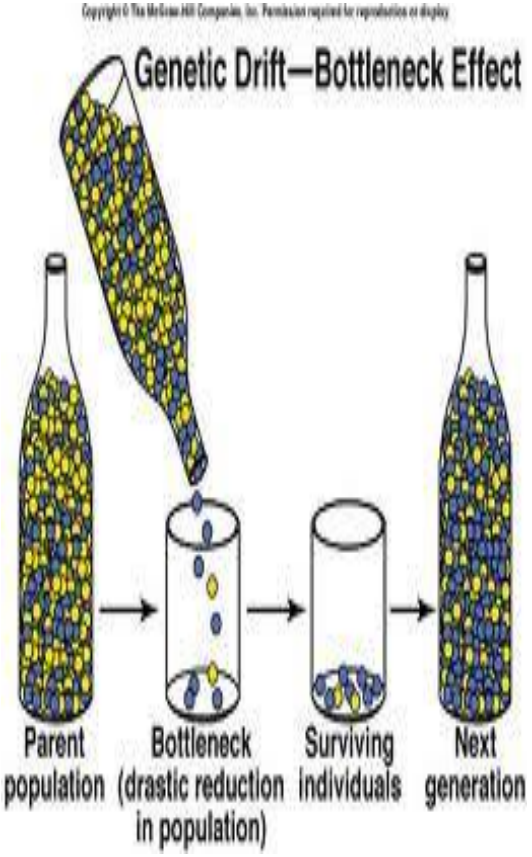
- The magnitude of random drift is directly measured by the ***effective population size***
- ***The effective population size of an actual population is the number of individual in a theoretical ideal population having the same magnitude of random genetic drift as the actual population***
- We aren't necessarily concerned about the census population size (N_c)
- We really want to know how many individuals are contributing to the next generation
- **Effective Population Size = N_e**

Not for distribution

SMALL POPULATION SIZE

- Small populations are characterized by drift and ***population bottlenecks***

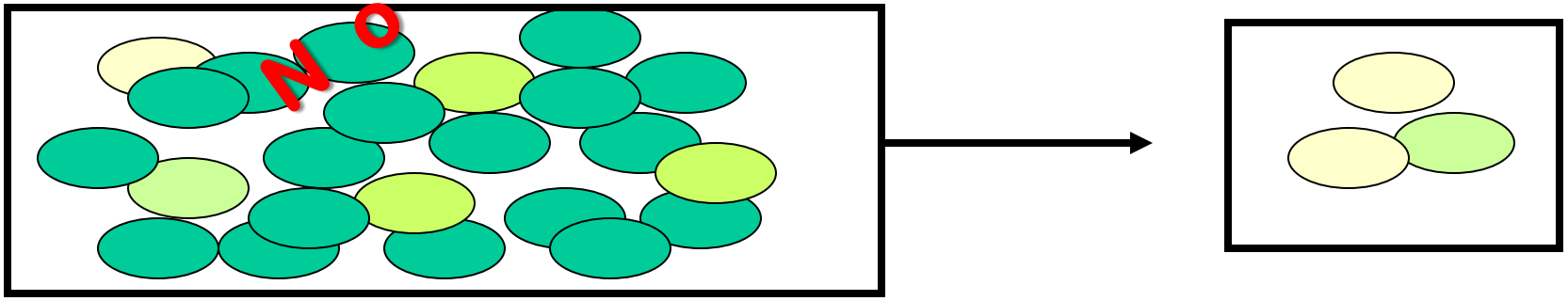




SMALL POPULATION SIZE

►New populations are sometimes associated with ***founder effects***: when a few individuals from a source populations found a new population, this population may not be representative of the genetic make-up of the original population. This is known as the ***founder effect***

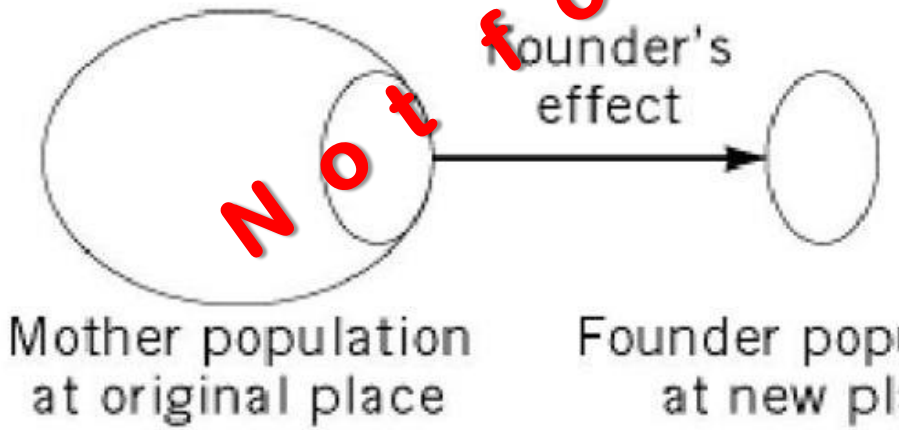
►There the founder effect is a form of drift. A severe population bottleneck (temporary reduction in size), which occurs in nature when a small group of emigrants from an established subpopulation founds a new population.

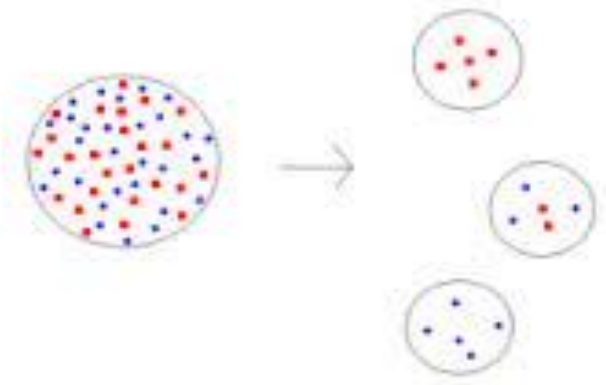


SMALL POPULATION SIZE

- ***Founder effect*** is a form of ***genetic drift*** that may occur in leading edge populations

- A severe ***population bottleneck*** (temporary reduction in size) often occurs in nature when a small group of emigrants from an established (sub)population founds a new population





EFFECTIVE POPULATION SIZE

- We aren't necessarily concerned about the census population size (N_c)
- We really want to know how many individuals are contributing to the next generation
- ***Effective Population Size = N_e***
- ***Ne is defined as the number of individuals that will contribute genes to the next generation by means of crossbreeding (Sewall Wright)***
- ***Ne*** *of an actual population is the number of individuals in a theoretical ideal population having the same magnitude of random genetic drift as the actual population*

EFFECTIVE POPULATION SIZE N_e

Since the effective population size is the crucial variable in determining the impact of drift, it is important to know the relationship between ***effective size (N_e) and census size (N_c)*** of a population

When effective population size is small (e.g., *N<50), then genetic drift becomes much more* important than selection (Motoo Kimura) and plays a paramount role in the evolutionary process (Doulgas Falconer)

Not for distribution

EFFECTIVE POPULATION SIZE N_e

- Since N_e can be half, much smaller, or by orders of magnitude less than N_c, under a number of different scenarios, it forms a very important genetic parameter (Russell Lande)

- N_e is a parameter in estimating genetic drift effects
  - Simple example: small European mountainous villages

N_e ESTIMATON

- There are several ways to estimate N_e
- **Sex ratio (M:F)**

$N_e = (4Nm \bullet Nf)/(Nm+Nf)$,

where Nm and Nf is the number of functional male and female in the population, respectively

500 M and 500 F; $N_e = 1000$

50 Males and 950 Females; $N_e = 190$

N_e ESTIMATION

- **Variance in family size**

$$Ne = 4Nc/(\sigma^2+2)$$

Variance in family size

Variance = 10, population of 120, Ne=40

N_e ESTIMATION

- **Fluctuation in population size between generations**
- A population is likely to vary in size with time. If N_i is the size at time *i* during a sequence of *t* generations, the effective size is the harmonic mean, and not the arithmetic mean of the sizes during this period

$$1/Ne=1/t(1/N_1+1/N_2+...+1/N_t)$$

Such a mean is strongly dependent on the smallest N_i in the sequence. Hence any generation of extremely ***small population size***, or ***bottleneck***, is very important in determining the amount of genetic drift in the whole sequences.

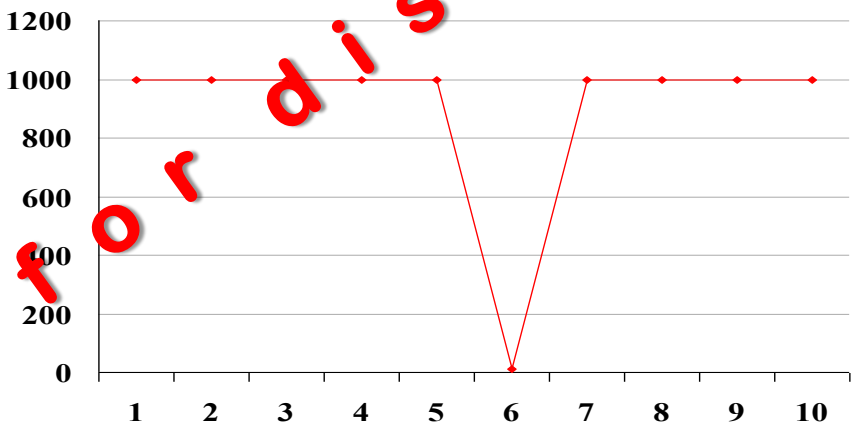
N_e ESTIMATION

- Fluctuations in population size

- Example (t in generations):

  N_c = 2000 9/10 years & 20 1/10years;

  N_e = 105.2



Not for distribution

Population Size

N_e ⟶

N_e ESTIMATION

- It is notoriously difficult to estimate *Ne in natural populations based on* ***demographic models,*** and currently the most widely used approaches employ ***genetic markers***.

- In addition, *genetic estimators* appear more conservative than *demographic models*.

- Coalescence theory proved useful in the prediction of *Ne,* in the evolutionary context for predicting genetic variability at the molecular level.

Erosion of heterozygosity

Another effect on genetic diversity through genetic drift is a decrease in heterozygosity

$$H_t = (1 - 1/N_e)^t \bullet H_o$$

Original Heterozygosity

INBREEDING

- When a population is small, especially if only a few individuals are reproducing, there is an increased likelihood of mating among close relatives, or **inbreeding**
- Inbreeding in a population acts to reduce the effective population size
  - N_e=N/(1+F), where F is the inbreeding coefficient
- Because individuals in outbreed populations tend to carry "lethal recessives" (lethal when homozygous), inbreeding can reduce fitness in a population, termed **inbreeding depression.**

Not for distribution