**Sampling genes: Drift, Identity By Descent, & Effective Population Size**

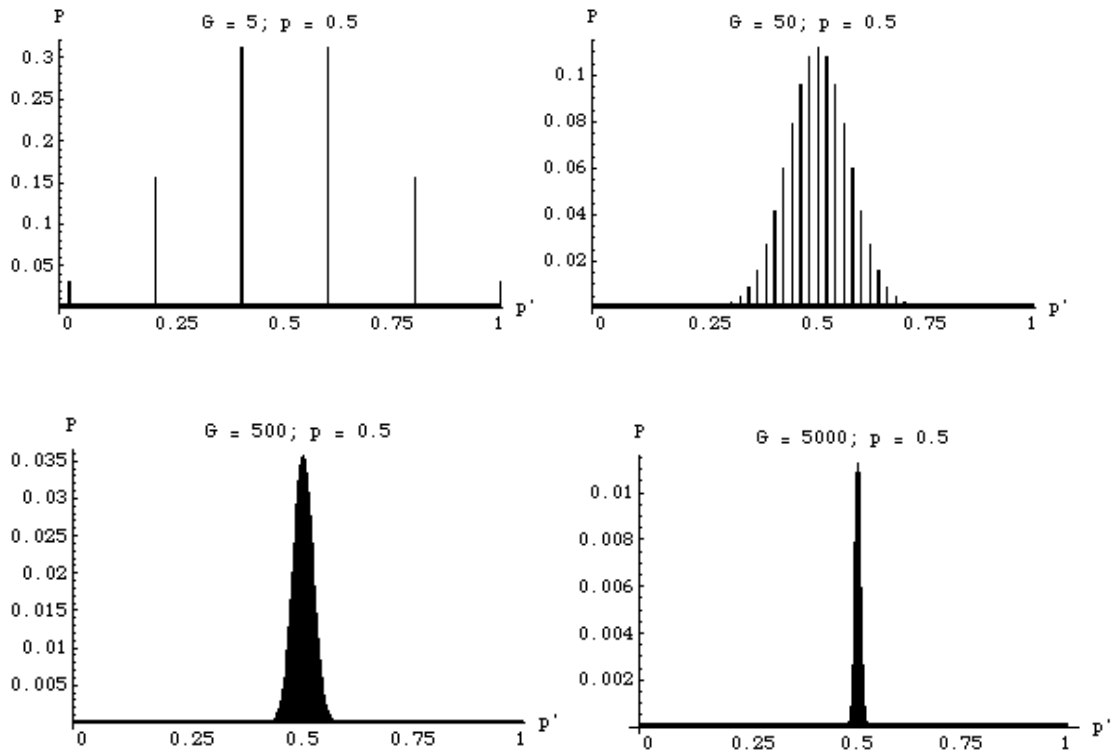
Imagine that a single male and female monkey (named Jack & Kate) are shipwrecked on a small island. Jack's genotype is *Aa* and Kate's genotype is *Aa*. Consequently, the frequency of *A* in this very small population is $p = ½$. Before they die, these two monkeys have two offspring. Both offspring happen to be genotype *AA*. The frequency of *A* is now $p = 1$.

**********
Q3.1) What is the probability of the two parental monkeys producing two offspring that are both *AA*?
Q3.2) In the scenario above, why did *p* change?
**********

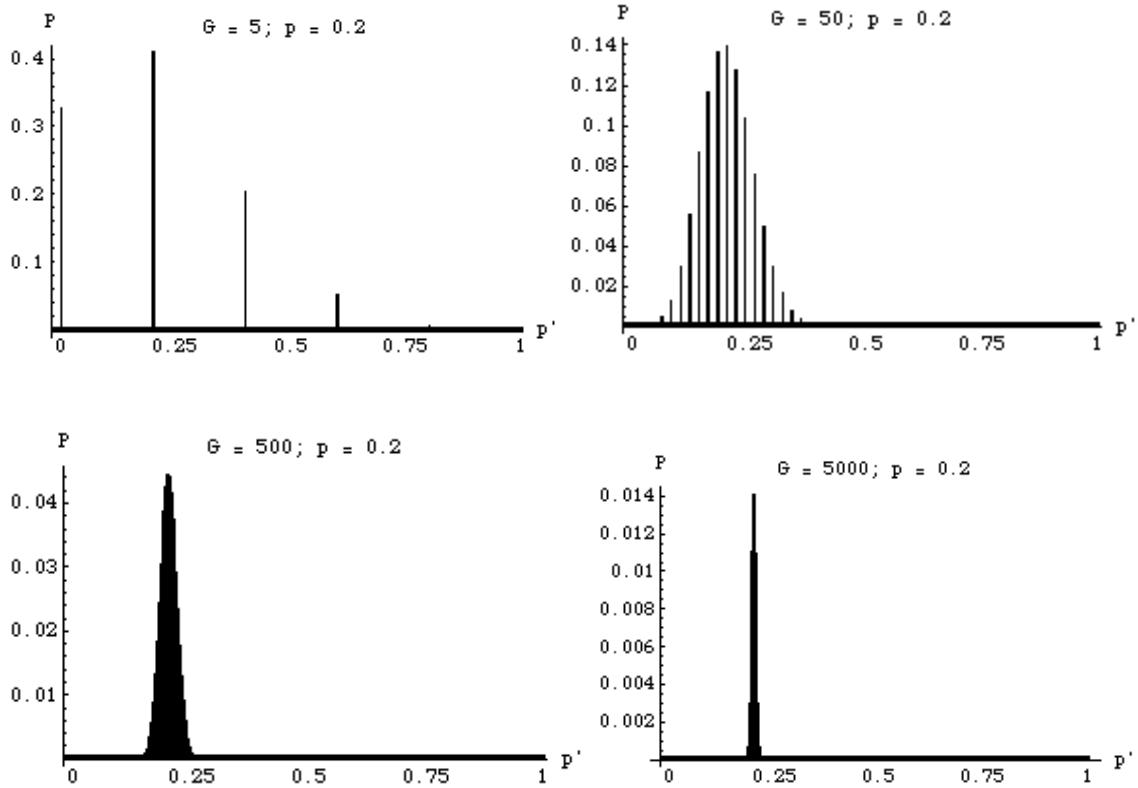
<u>Genetic drift</u> refers to changes in allele frequencies due to the sampling of gametes. Consider a population in which the frequency of the *A* allele is *p*. Under random mating, you can think reproduction as randomly choosing alleles out of a gamete pool. We assume that a very large number of gametes is produced so that the frequency of *A* in the gamete pool remains *p* regardless of which gametes are chosen. To produce a population of *N* diploid individuals, 2*N* gametes must be chosen. The probability of choosing *x* *A* alleles is given by the binomial probability distribution:

$$P(x) = \binom{2N}{x} p^x (1-p)^{2N-x}$$

The figures below show the probability, *P,* that a sample of *G* gametes will have a frequency of *A* equal to *p′*.



As above, the figures below show the probability, *P,* that a sample of *G* gametes will have a frequency of *A* equal to *p′*. However, this time the initial frequency of *A* is $p = 0.2$.




The figures above illustrate that when the sample of gametes is small, there is a fairly good chance that *p′* will be quite different than *p*. When the sample size is large, *p′* will be very close to *p*.

**Changes in the mean and variance.**
The expected value is $E(x) = 2Np$. Thus, the expected frequency of *A* in the next generation is $E(p') = E(x)/(2N) = p$. This means that the expected change due to sampling is $E(\Delta p) = E(p' - p) = E(p') - E(p) = 0$. In other words, *on average*, drift causes no change in allele frequency.

Let us consider the variance in *p′*. What is the variance in the number of *A* alleles sampled? $V(x) = 2Np(1 - p)$. Because $p' = x/(2N)$, we know that $V(p') = p(1 - p)/(2N)$ and therefore $V(\Delta p) = V(p' - p) = V(p') = p(1 - p)/(2N)$. In other words, increasing the population size reduces the variation in allele frequency change due to drift.

Note that there can be no drift if $p = 0$ or $p = 1$. If a population gets to 0 or 1 by chance, it will be stuck there (assuming no other forces like mutation or migration). Because of sampling, allele frequencies vary over time and so eventually the population will hit one of these *absorbing boundaries*.

Imagine that a species is divided into a large number of small subpopulations (demes). We can decompose the total variance ($V_T$) into two parts: variance *within* demes ($V_W$) and variance *among* demes ($V_A$). The former being related to the differences among individuals from the same deme and the latter being related to the differences between the averages of different demes. If the demes are initially identical (i.e., $V_A = 0$; $V_W = V_T$) and are allowed to drift indefinitely with no migration, different populations will become fixed for different alleles. This means that at some point in the future, there will be no variance within populations ($V_W = 0$) but there will be variance among populations ($V_A = V_T$). This is why drift is said to cause a conversion of variance within populations to variance among populations.

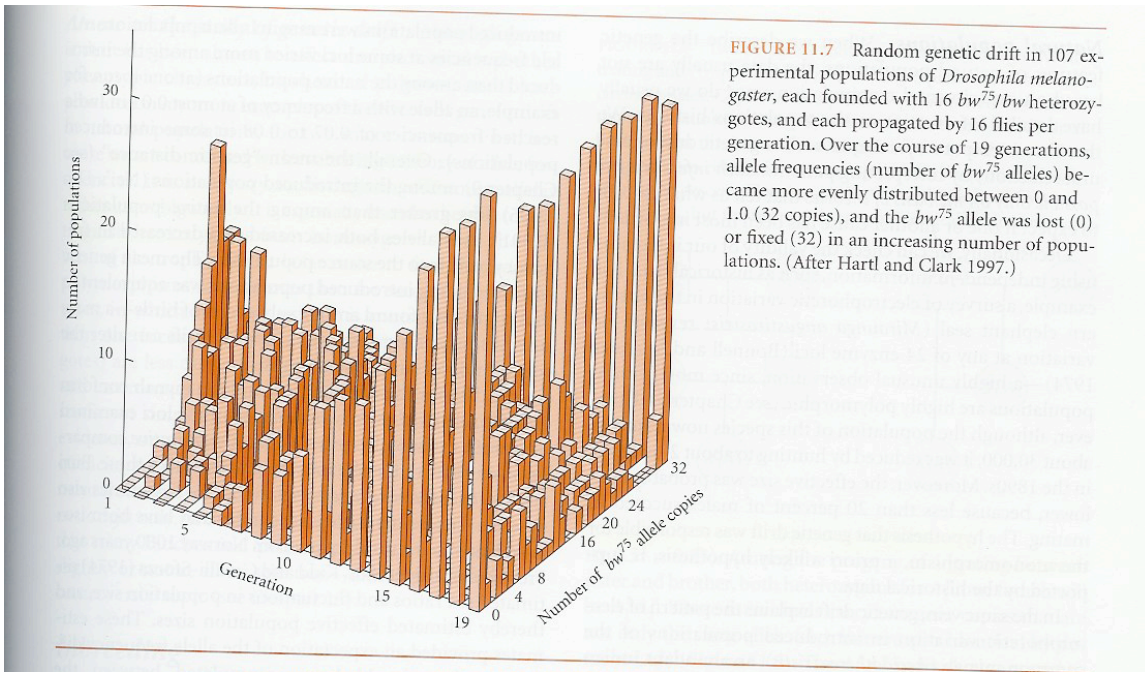



FIGURE 11.7 Random genetic drift in 107 experimental populations of *Drosophila melanogaster*, each founded with 16 $bw^{75}/bw$ heterozygotes, and each propagated by 16 flies per generation. Over the course of 19 generations, allele frequencies (number of $bw^{75}$ alleles) became more evenly distributed between 0 and 1.0 (32 copies), and the $bw^{75}$ allele was lost (0) or fixed (32) in an increasing number of populations. (After Hartl and Clark 1997.)

The figure above (from the Futuyma textbook, p. 305) illustrates the results of an experiment by Peter Buri (1956). Buri initiated 107 lines each with 16 heterozygous flies ($p$ =0.5, $N$ =16). Each generation, eight flies of each sex were selected at random to produce the next generation. After 19 generations, many of the lines are fixed for one of the two alleles ($p$ = 0 or $p$ = 1).

## Identity by descent (Inbreeding)

Consider a constant population of $N$ diploid hermaphrodites that mate at random (including selfing). Imagine further that each of the $2N$ alleles at one neutral locus is unique at time $t = 0$, where time is measured in generations. At the next generation, $t = 1$, there will also be $N$ individuals with $2N$ alleles but, owing to sampling, some alleles present at $t = 0$ will have been lost and some alleles present at $t = 0$ will exist in multiple

copies; they will no longer be unique. As a result, the genetic variance in this descendant population will be less than that of our starting generation.

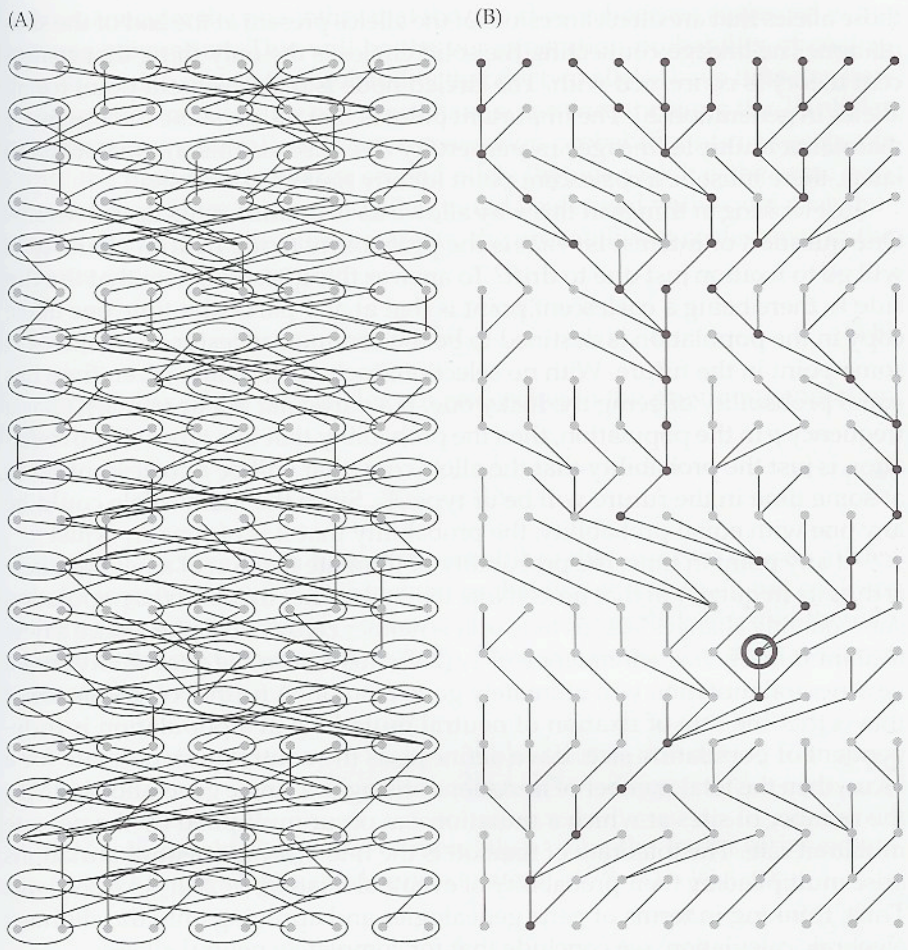



**Figure 3.4** Simulation of the gene genealogy for a population of $2N = 10$ allele copies over 20 generations. Time proceeds from bottom to top. Each generation is produced by choosing an allele copy at random from the previous generation to be the parent for one of the copies in the next generation. The ovals in (A) represent diploid individuals. In (B), the allele copies within each generation after the first are rearranged in order to untangle the lineages. The black circles are those copies that are ancestors of all of the allele copies in generation 20. The circled node is the coalescent point for all alleles in generation 20.

*Figure from Rice (2004).*

How do we describe this process of loss of genetic variation by random genetic drift, the change in gene frequency that occurs in finite breeding populations as a result of sampling?

Define the probability of "***identity by descent***" (**IBD**) as $F(t)$, the probability that two alleles at generation $t$ are descended from the same parent allele in generation $t = 0$. Thus our starting point, where every allele is unique, becomes our reference point for determining **IBD**. Furthermore, $F(0) = 0$ by definition of "unique" allele.

We assume that a very large (infinite) number of gametes are produced and each allele exists at a frequency of $1/2N$.

The probability that two alleles in generation 1 are both direct descendents of the same allele from generation 0 is

$$F(1) = 1/2N \qquad [3.1]$$

(i.e., sampling at random and with replacement from the $2N$ alleles at $t = 0$)

Let us consider what $F(2)$ will be. There are now two ways that alleles could be IBD.

$$\begin{aligned} F(2) &= (\text{prob. of IBD from } t = 1) + (\text{not IBD from } t = 1)(\text{IBD prior to } t = 1) \\ &= 1/2N + (1 - 1/2N)F(1) \\ &= 1/2N + (1 - 1/2N)(1/2N) \\ &= (1/2N)(1 + (1 - 1/2N)) \qquad [3.2] \end{aligned}$$

In words, this equation says that the probability that two alleles sampled at generation 2 are IBD is equal to the probability that they are both direct descendants of the same allele from the preceding generation ($1/2N$) plus the probability that they descended from different alleles from the preceding generation ($1 - 1/2N$) but that these two alleles in the previous generation were IBD.

In the third generation,

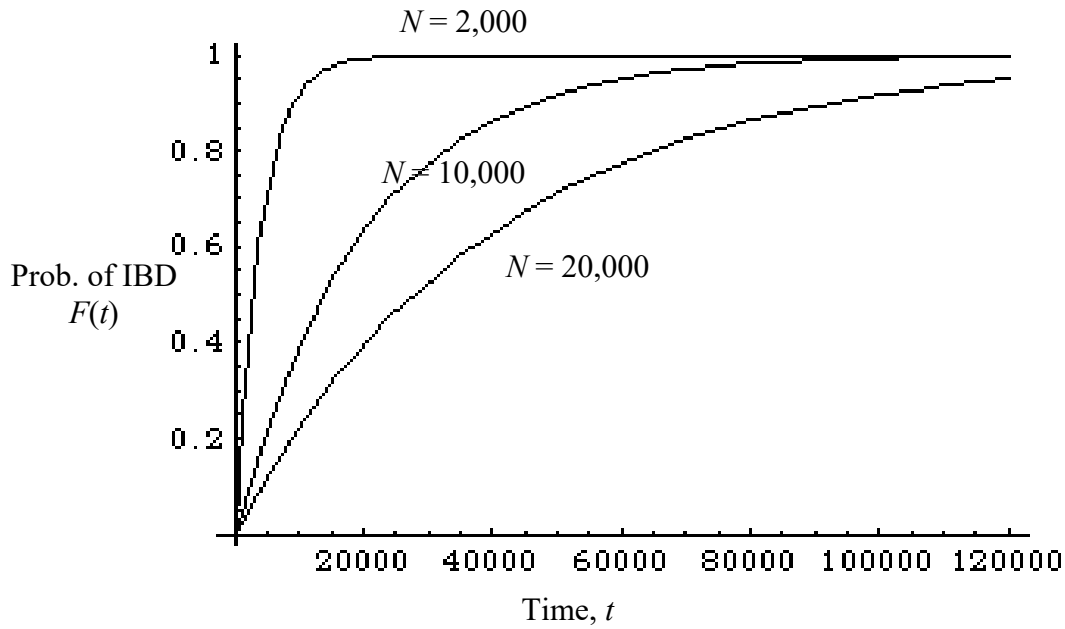
$$\begin{aligned} F(3) &= (\text{prob. of IBD from } t = 2) + (\text{not IBD from } t = 2)(\text{IBD prior to } t = 2) \\ &= 1/2N + (1 - 1/2N)F(2) \\ &= 1/2N + (1 - 1/2N)\,(1/2N)(1 + (1 - 1/2N)) \\ &= 1/2N + (1/2N)((1 - 1/2N) + (1 - 1/2N)^2) \\ &= (1/2N)\,(1 + (1 - 1/2N) + (1 - 1/2N)^2) \qquad [3.3] \end{aligned}$$

And in the following generation,

$$\begin{aligned} F(4) &= (\text{prob. of IBD from } t = 3) + (\text{not IBD from } t = 3)(\text{ IBD prior to } t = 3) \\ &= 1/2N + (1 - 1/2N)F(3) \\ &= 1/2N + (1 - 1/2N)\,(1/2N)\,(1 + (1 - 1/2N) + (1 - 1/2N)^2) \\ &= 1/2N + (1/2N)((1 - 1/2N) + (1 - 1/2N)^2 + (1 - 1/2N)^3) \\ &= (1/2N)\,(1 + (1 - 1/2N) + (1 - 1/2N)^2 + (1 - 1/2N)^3) \qquad [3.4] \end{aligned}$$

In general,

$$\begin{aligned} F(t) &= (\text{prob. of IBD from } t - 1) + (\text{not IBD from } t - 1)(\text{IBD prior to } t - 1) \\ &= 1/2N + (1 - 1/2N)F(t - 1) \qquad [3.5a] \\ &= \frac{1}{2N}\sum_{k=0}^{t-1}\left(1 - \frac{1}{2N}\right)^k = 1 - \left(1 - \frac{1}{2N}\right)^t \qquad [3.5b] \end{aligned}$$




*The figure above shows us two important things about the probability that two alleles will be identical by descent F(t).*

*(1) In the absence of any other forces (e.g. mutation), the probability increases until it reaches 1.*

*(2) F(t) increases faster for smaller populations than it does for larger populations.*

Let us consider what happens after some long period of time, i.e., $t \to \infty$. Note that the term in brackets in [3.5b] is between 0 and 1, i.e., $0 < (1 - 1/2N) < 1$. For any number $x$ where $0 < x < 1$, $x^t \to 0$ as $t \to \infty$. Therefore the equation above

$$F(\infty) = 1 - \left(1 - \frac{1}{2N}\right)^{\infty} = 1 - 0 = 1. \qquad [3.6]$$

This result means that eventually all alleles in the population will be identical by descent eventually (i.e., all descended from a single ancestral allele). Unless some force introduces new variation, genetic drift will eliminate all variation within this population.

Heterozygosity $H(t)$ can be defined as the probability that two randomly sampled alleles will not be IBD, i.e., $H(t) = 1 - F(t)$. We can re-write this as $F(t) = 1 - H(t)$. Substituting into [3.5a] we obtain

$$F(t) = 1/2N + (1 - 1/2N)\, F(t - 1)$$
$$1 - H(t) = 1/2N + (1 - 1/2N)\, [1 - H(t\text{-}1)]$$
$$-H(t) = 1/2N - 1/2N - (1 - 1/2N)\, H(t\text{-}1)$$
$$H(t) = (1 - 1/2N)\, H(t\text{-}1)$$

Remembering that the equation above implies that $H(t\text{-}1) = (1 - 1/2N)\, H(t\text{-}2)$, we can write the equation above as

$$H(t) = (1 - 1/2N)\, H(t\text{-}1) = (1 - 1/2N)\,(1 - 1/2N)\, H(t\text{-}2) = (1 - 1/2N)^2\, H(t\text{-}2)$$

Following this logic, we could use $H(t\text{-}2) = (1 - 1/2N)\, H(t\text{-}3)$ and write the equation as

$$H(t) = (1 - 1/2N)^2\, H(t\text{-}2) = (1 - 1/2N)^2\,(1 - 1/2N)\, H(t\text{-}3) = (1 - 1/2N)^3\, H(t\text{-}3)$$

You should notice a pattern. Continuing with this logic back $t$ generations, we'd find that

$$H(t) = (1 - 1/2N)^t\, H(t\text{-}1) = (1 - 1/2N)^t\, H(0)$$

**********

Q3.3) Sketch a graph of $H(t)$ over time for a large population and a small population assuming $H(0) = 0.7$.

**********

<u>Drift-Mutation Balance (infinite alleles)</u>

If there is mutation or migration, then the loss of variation by drift can be balanced such that an equilibrium level of IBD is reached. Now lets add mutation to the model. We will assume each generation an allele mutates with probability $\mu$. Each mutation is assumed to be unique (i.e., new mutations are not IBD with any other allele – the infinite alleles model).

The probability that two alleles are IBD at generation $t$ is

$$F(t) = (1-\mu)^2 \frac{1}{2N} + (1-\mu)^2\left(1 - \frac{1}{2N}\right)F(t-1) \qquad [3.7]$$

**********

Q3.4) Explain equation [3.7] in words. Hint: note its similarities and differences to [3.5].

**********

Expanding [3.7], we have

$$F(t) = \frac{1-F(t-1)}{2N} + (1-2\mu)F(t-1) + \mu^2\left(F(t-1)\left(1-\frac{1}{2N}\right) + \frac{1}{2N}\right) + \frac{\mu}{2N}\left(1-F(t-1)\right) \qquad [3.8]$$

If we assume that $\mu$, $1/2N << 1$, then terms involving $\mu^2$ or $\mu/2N$ will be very small relative to the other terms, so we can ignore them,

$$F(t) \approx \frac{1-F(t-1)}{2N} + (1-2\mu)F(t-1) \qquad [3.9]$$

At equilibrium, $F(t) = F(t-1) = F^*$

$$F^* \approx \frac{1-F^*}{2N} + (1-2\mu)F^* \qquad [3.10]$$

Solving for $F^*$ we find,

$$F^* \approx \frac{1}{1+4N\mu} \qquad [3.11]$$

As we will see, $4N\mu$ arises so frequently in population genetics it has been given its own symbol, $\theta = 4N\mu$.

**********
Q3.5) At mutation-drift equilibrium, what is the probability an individual will be heterozygous? (I.e., what is the probability that two alleles will be different?)
**********

<u>Drift-Migration Balance</u>
Imagine instead that you are sampling alleles from a single subpopulation or deme within a metapopulation. The primary source of "new" alleles into the population would be from migration rather than mutation, assuming $m >> \mu$ where $m$ is the migration rate. The analysis proceeds exactly as above using $m$ in place of $\mu$ (we must still assume that $m << 1$). Because we are sampling alleles from a single subpopulation we can think of $F$ as $F_{ST}$ and we obtain the classic island model result

$$F_{ST}^* \approx \frac{1}{1+4Nm}$$

Numerous people have used this result to infer the migration rate $m$ because they are able to estimate $F_{ST}$ from molecular data and $N$ from ecological data. However, the equation above is based on the "island model" which assumes that each subpopulation is small relative to the total metapopulation and that each population makes an equal (and very small) contribution to the migrant pool so that alleles migrating into the focal population are unique. Most real populations do not meet the assumptions of the island model so it can be very misleading to attempt to estimate $m$ using the equation above (see Whitlock and McCauley 1999 for discussion of this issue).

**Fixation of neutral alleles.**

In the absence of mutation or migration, the probability of IBD goes to 1 reflecting the fact that one lucky allele goes to fixation, while all others go extinct. Because each of the $2N$ alleles in a population is equally likely to the one that goes to fixation, it is clear the probability of fixation for any given single copy of an allele is $1/2N$. If there are $x$ copies of a certain type of allele, each copy will have a probability $1/2N$, the probability that one of these copies goes to fixation is simply $x/2N$. That is the probability that an allele drifts to fixation is equal to its frequency.

If two populations (one of size $N_1$ and the other of size $N_2$), are isolated for a long time, say $t$ generations. The level of sequence divergence between them is expected to be $2\mu t$, where $\mu$ is the mutation rate.

**********
Q3.6) Explain why. Hint: What is the frequency of a new mutation and so what is the probability this mutation will become fixed? How many new mutations enter the population each generation? Can you think of what assumptions we have made in doing this calculation.
**********

**Effective population size: Quantifying the strength of drift through a summary measure**

When we think about selection, we might want to compare the strength of selection acting on the same site in different species or at different sites across the genome within a species. Under a standard fitness model, we might do this by comparing $s$ values. In the same sense, we'd like to compare the strength of genetic drift experienced by the same site in different species or at different sites across the genome within a species. How can we do this?

So far, we have been considering what is known as the "ideal" population. By "ideal", we mean that it is the mathematically simplest possible scenario: (i) a population of constant size, (i) each generation every individual makes an equal and very large (effectively infinite) number of gametes, and (iii) these gametes are sampled at random (with no constraints) to produce the next generation. Under these ideal conditions, we can quantify the effects of gamete sampling (i.e., drift) in a reasonably simple and straightforward manner as we have done above. No matter what aspect of drift we are looking at (e.g., variance in the change in allele frequency, rate of inbreeding), the magnitude of the drift effect depends on only one parameter, the population size $N$. (Of course, this is because there is only one parameter under the assumptions of the ideal population!) In other words, the strength of drift is captured entirely by $N$ in the ideal population.

Sampling effects in real populations are more complicated because real populations violate the assumptions of the ideal population in numerous ways: (i) populations change in size over time; (ii) some individuals produce more gametes than others for both genetic and non-genetic reasons; and (iii) gametes are not sampled completely at random (e.g., due to the requirements of separate sexes in some species, because of spatial structure, etc.). Different species violate the "ideal" assumptions in different ways or to different extents. Different sites within the genome also vary in how much drift they experience because of these violations (especially with respect to (ii)). The strength of drift depends, sometimes in complicated ways, on these various violations. So how do we quantify drift using a common currency that allows us to compare the strength of drift across species or across sites? In population genetics, we use drift in the ideal population as the standard for quantification and comparison. Loosely speaking, we can think of it as follows. We convert the 'amount' of drift experienced by our actual site of interest to the amount of

drift experienced in an ideal population. If our actual site experiences an amount $x$ of drift, we ask what population size would an ideal population have to be to experience this same amount, $x$, of drift. We call this size the effective population size, $N_e$. With respect to drift, our actual site is *effectively* like a site in an ideal population of size $N_e$. Many people think of $N_e$ as quantifying something about the number of individuals in the population. It is more helpful to think of $N_e$ as the currency for quantifying the amount of drift (which does depend on population size but also other things too).

Here's a bit of a silly (and imperfect but possibly helpful) analogy. Imagine there was an island where the people only burned wood to produce heat. You could quantify each individual's capacity to produce heat entirely by the kilograms of wood that person possessed, $W$. Now consider people in other parts of the world. A person, Bob, might have very little wood but has some gas, oil, and coal. One could think of Bob's heating making capacity in terms of how much wood someone would need to produce the same amount of heat as Bob could with his supply of wood, gas, oil, and coal, i.e., Bob "effectively" has that much wood. Bob's effective wood value, $W_e$, does not really tell us anything about how much wood Bob has. It is just saying that with respect to heat production, Bob is like a person with $W_e$ amount of wood. Note, that gas, oil, and coal differ from wood in various respects (e.g., the rate at which heat can be produced, maximum temperature). Thus, converting Bob's assets into "effective amount of wood" does not capture all aspects of heat production; analogously $N_e$ population size does not capture all aspects of genetic drift.

Different types of effective population sizes

We have seen that sampling affects the variance in the change in allele frequency as well as the rate of inbreeding. So far, we have been considering an 'ideal' population, i.e., we assumed a constant population size and completely random mating and many other things. The $Q$ effective population size, ${}_QN_e$, represents the size of an ideal population that would affect some drift-related property $Q$ in the same way as non-ideal population of size $N$. For example, the *inbreeding* effective population size, ${}_IN_e$, represents the size of an ideal population that would have a rate of inbreeding equivalent to a non-ideal population of size $N$.

Inbreeding Effective Population Size, ${}_IN_e$

*Fluctuating Population Size.*

Imagine that a population changes in size every generation such that the population size in generation $x$ is denoted as $N_x$ (e.g., $N_0$, $N_1$,… $N_t$). What would be the population size of an ideal population that had the same rate of inbreeding (i.e., what is ${}_IN_e$)?

Recall that the probability that two alleles are not IBD after $t$ generations in an ideal population is

$$1 - F(t) = (1 - 1/(2N))^t \cong e^{-t/(2N)}$$

NOTE: $e^{-x} \cong (1 - x)$ for $|x| << 1$ (a 1st order Taylor series approximation)

If the population size changes each generation then,

$$1 - F(t) = (1 - 1/(2N_0))\,(1 - 1/(2N_1))\ldots(1 - 1/(2N_{t-1})) \cong$$
$$e^{-1/(2N_0)}e^{-1/(2N_1)}\ldots e^{-1/(2N_{t-1})} = e^{-1/(2N_0)-1/(2N_1)\ldots-1/(2N_{t-1})}$$

Now, let us set the equation using the ideal (constant) population size equal to the equation using the variable population size and solve for ${}_IN_e$

$$e^{-t/(2\,{}_IN_e)} = e^{-1/(2N_0)-1/(2N_1)\ldots-1/(2N_{t-1})}$$

$$t/(2\,{}_IN_e) = 1/(2N_0)) + 1/(2N_1)\ldots 1/(2N_{t-1})$$

$$\frac{1}{{}_IN_e} = \frac{1}{t}\sum_{j=0}^{t-1}\frac{1}{N_j}$$

The effective population size is equal to the harmonic mean population size. The harmonic mean is dominated by small numbers. If the population size over six generations is $10^2$, $10^3$, $10^4$, $10^5$, and $10^6$, the effective population size is 450 even though the arithmetic average is 222,220.

*Breeding Sex Ratio.*
In many species, mating is not random in that it must occur between males and females. For autosomal loci, 50% of alleles are transmitted by females and 50% by males.

The probability that two randomly samples alleles will be *recently* IBD is equal to the probability that both alleles come from the same parent (which is either male *or* female) and are the same allele within that parent (1/2).

$P_{rIBD}$ = *P(both alleles from a female) P*(*same female*) *P*(same allele within that female)
+ *P(both alleles from a male) P*(*same male*) *P*(same allele within that male)

$$= (1/4)(1/(2N_{f,t-1})) + (1/4)(1/(2N_{m,t-1}))$$

$$= 1/(8N_{f,t-1}) + 1/(8N_{m,t-1})$$

In an ideal population,

$$P_{rIBD} = 1/(2\,{}_IN_e)$$

Setting $1/(2\,{}_IN_e) = 1/(8N_{f,t-1}) + 1/(8N_{m,t-1})$, we find that

${}_IN_e = 4N_{f,t\text{-}1}N_{m,t\text{-}1}/(N_{f,t\text{-}1} + N_{m,t\text{-}1})$

Note that when there is an even sex-ratio, $N_{f,t\text{-}1} = N_{m,t\text{-}1} = N_{T,t\text{-}1}/2$, then ${}_IN_e = N_{T,t\text{-}1}$.

**********
Q3.7) The result above indicates that variation in the sex ratio can alter the rate of inbreeding. Consider a case where there are nine females for every male. What is the effective population size if the total population size is $N_T$.
**********

*Variation in family size.*
Our previous results have made a hidden assumption about variation in family size. We can see the importance of variation in family size if we make a more detailed analysis (following the approach used by S. Rice in his 2004 book).
Assume $N_{t\text{-}1}$ parents produce $N_t$ offspring.
Let $k_i$ be the number of successful gametes (and offspring – no selfing) from parent $i$.
We first need to know probability that two alleles came from the same parent.
The probability that two alleles came from the $i$th parent is

$$S_i = \text{A*B*C} = \left(\frac{N_t - 1}{N_t}\right)\left(\frac{k_i}{2N_t}\right)\left(\frac{k_i - 1}{2N_t - 2}\right)$$

where
A is the probability of selecting alleles from two different offspring, A $=(N_t - 1)/N_t$
B is the probability that the first allele came from parent $i$, B $= k_i/(2N_t)$
C is the probability that the second allele came from parent, C $= (k_i - 1)/(2N_t - 2)$

**********
Q3.8) Explain in words why C $= (k_i - 1)/(2N_t - 2)$.
**********

We want to know the probability that two alleles came from the same gene copy in the previous generation, $P_{rIBD}$.

$$P_{rIBD} = \sum_{i=1}^{N_{t-1}} \frac{1}{2} S_i$$

**********
Q3.9) Explain why this equation makes sense.
**********

$$P_{rIBD} = \sum_{i=1}^{N_{t-1}} \frac{1}{2} S_i = \frac{1}{2} \sum_{i=1}^{N_{t-1}} \left( \frac{k_i}{2N_t} \frac{(k_i - 1)}{2N_t} \right)$$

**********
Q3.10) Show the intermediate steps needed to get to the right-hand side.

**********

Note that $2N_t = \bar{k}N_{t-1}$

**********

Q3.11) Explain why.
**********

$$P_{rIBD} = \frac{1}{2}\sum_{i=1}^{N_{t-1}}\left(\frac{k_i}{\bar{k}N_{t-1}}\frac{(k_i-1)}{\bar{k}N_{t-1}}\right) = \frac{1}{2\bar{k}^2N_{t-1}}\left(\frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}}k_i^2 - \frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}}k_i\right)$$

On the right-hand side, the first term in parentheses is $E(k^2)$ and second term is $E(k)=\bar{k}$. Recalling the $V(k) = E(k^2) - \bar{k}^2$, we get

$$P_{rIBD} = \frac{1}{2\bar{k}^2N_{t-1}}(V(k) + \bar{k}^2 - \bar{k})$$

Recalling that in an ideal population, $P_{rIBD} = 1/(2\,{}_IN_e)$, we get

$${}_IN_e = \frac{\bar{k}^2N_{t-1}}{V(k) + \bar{k}^2 - \bar{k}}$$

Assuming the population size is constant, $\bar{k} = 2$, then

$${}_IN_e = \frac{4N_{t-1}}{V(k) + 2}$$

Assuming $V(k) = \bar{k} = 2$, then ${}_IN_e = N_{t-1}$. In other words, we had originally been implicitly making this assumption. In our ideal population, we had been implicitly assuming that the distribution of family sizes was Poisson distributed. In the Poisson distribution, $V(k) = \bar{k}$.

**********

Q3.12) Assuming the population cannot grow in size, how could a conservation biologist managing a small breeding population of *Mustela nigripes* use this result to reduce the rate of loss of variation?
**********

Variance Effective Population Size, ${}_VN_e$

The variance effective population size is the population size of an ideal population that would cause an equal amount of variance in the change in allele frequency as a real population. Recall that for an ideal population, $V(\Delta p) = p(1-p)/(2\,{}_vN_e)$.

*Breeding Sex Ratio*

Let the frequency of the $A$ allele be $p$ in the gamete pool the produces the current generation. Of the offspring produced, $N_m$ males survive to reproduce and $N_f$ females. Because males and females contribute equally to following generation, $p' = (p_f' + p_m')/2$, where $p_f'$ is the frequency of $A$ among the $N_f$ females. It can be shown that,

$$V(\Delta p) = \frac{p(1-p)}{8}\left(\frac{1}{N_m} + \frac{1}{N_f}\right)$$

**********

Q3.13) Show how to get this.

**********

Consequently,

$${}_VN_e = \frac{4N_fN_m}{N_f + N_m}$$

**********

Q3.14) Show how to get this.

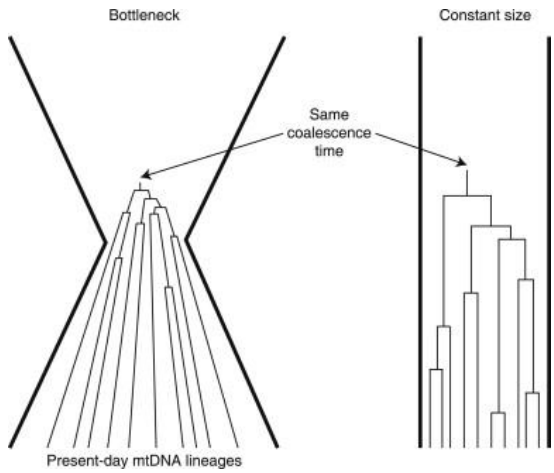
**********

Note this result is very similar to the inbreeding effective size result for different numbers of males and females, but there is a slight difference. The variance result depends on the numbers of males and females in the *current* generation whereas the inbreeding result depends on these numbers in the previous generation. This makes sense because inbreeding depends on the probability that two alleles came from the same gene copy and this will depend on the number of parents. In contrast, variance depends on the number of gametes being chosen to produce the current generation and this is determined by the number of individuals in the current generation.

**A note of caution on ${}_QN_e$ (effective population size isn't everything!)**

The concept of effective population size is very useful as it provides a "summary" of how *some aspect* of drift works in "non-ideal" cases. Often, we can then use other results developed in the ideal case by simply replacing $N$ with $N_e$ and this will give us a good approximation of how things work in the non-ideal case. However, ${}_QN_e$ only gives equivalence with respect to some particular aspect of drift and this does not extend to all aspects of drift equally well. One example comes when we think about genealogies. Fluctuations in population size not only affect the total length of a multi-sample

genealogy but also its shape. Changes in the shape of the genealogy will cause *Tajima's D* to deviate from zero (we expect *Tajima's D* ≈ 0 the ideal case for any *N*) so a simple substitution of $N_e$ for *N* will not suffice to capture this aspect of fluctuating population size.




Modified figure from Weaver & Roseman (2008). Note the excess of long terminal branches in the bottlenecked population on the right.