Agronomy/Animal Science 561
Quantitative and Population Genetics for Breeding
Fall 2000

Contents

1 Background

1.1 Mendel's Laws

- *The law of segregation.* A trait is determined by pairs of factors, but gametes contain only one of these chosen at random.
- *The law of independent assortment.* Factors from parents combine independently in offspring.

2 Basic Concepts in Probability and Statistics

2.1 Random Variable

Definition 1 *When the value of a variable, Y, is determined by some random process, Y is called a random variable.*

Example 1 *Suppose the height, Y, of a plant is 100 units when the genotype at a locus A is AA or Aa, and is 50 units when the genotype is aa. Then, the height of a randomly sampled plant is a random variable.*

2.2 Sample Space

Definition 2 *The set of possible values for a random variable is called the sample space of the random variable.*

Example 2 *The random variable in example (1), has a sample space of (50,100).*

2.3 Probability (by example)

Example 3 *Consider determining the genotype, T, for each of N randomly sampled plants.*
N_{AA} *= plants with genotype AA*
N_{Aa} *= plants with genotype Aa*
N_{aa} *= plants with genotype aa*
As N becomes very large,

$$\Pr(T = AA) = \frac{N_{AA}}{N}$$

$$\Pr(T = Aa) = \frac{N_{Aa}}{N}$$

$$\Pr(T = aa) = \frac{N_{aa}}{N}$$

Example 4 *Suppose two coins are flipped N times. After each flip the number of heads, Y, is determined*
N_0 = *number of times* $Y = 0$
N_1 = *number of times* $Y = 1$
N_2 = *number of times* $Y = 2$
As N becomes very large,

$$\Pr(Y = 0) = \frac{N_0}{N}$$

$$\Pr(Y = 1) = \frac{N_1}{N}$$

$$\Pr(Y = 2) = \frac{N_2}{N}$$

2.4 Expected Value

Definition 3 *Let the sample space for random variable Y be denoted by $y_1, y_2, \ldots, y_k$ and let*

$$\Pr(y_i) = \Pr(Y = y_i)$$

Then, the expected value of Y is defined as

$$E(Y) = \sum_{i=1}^{k} y_i \Pr(y_i)$$

The expected value is a measure of the location of the distribution.

Example 5 *Suppose the 305 day milk yield, Y, in cows is related to the genotype at locus A as follows:*

Genotype	*Probability*	*Milk Yield (arbitrary units)*
aa	*0.2*	*100*
Aa	*0.5*	*150*
AA	*0.3*	*200*

Then, the expected value of Y is

$$\begin{aligned} E(Y) &= 100 \times \Pr(Y = 100) + 150 \times \Pr(Y = 150) + 200 \times \Pr(Y = 200) \\ &= 100(0.2) + 150(0.5) + 200(0.3) \\ &= 155 \end{aligned}$$

2.5 Variance

Definition 4 *The variance of a random variable Y is defined as*

$$Var(Y) = E\{[Y - E(Y)]^2\}$$

The above can also be written as

$$Var(Y) = E(Y^2) - [E(Y)]^2$$

The variance is a measure of the spread of the distribution.

Example 6 *Consider computing the variance for milk yield given in example (5). Using the definition of expected value,*

$$\begin{aligned} E(Y^2) &= (100)^2(0.2) + (150)^2(0.5) + (200)^2(0.3) \\ &= 25250 \end{aligned}$$

From example (5), $E(Y) = 155$, so

$$\begin{aligned} Var(Y) &= 25250 - (155)^2 \\ &= 1225 \end{aligned}$$

2.6 Joint Probability

Definition 5 *The probability of two or more random variables.*

Example 7

Joint Probabilities for Genotype and Milk Yield

	Milk Yield (arb. units)		
Genotype	100	200	300
aa	0.175	0.05	0.025
Aa	0.1	0.3	0.1
AA	0.025	0.05	0.175

2.7 Conditional Probability

Definition 6

$$Pr(X = x|Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$$

Example 8

Conditional Probabilities for Milk Yield given Genotype

	Milk Yield (arb. units)		
Genotype	100	200	300
aa	0.7	0.2	0.1
Aa	0.2	0.6	0.2
AA	0.1	0.2	0.7

2.8 Conditional Expectation

Definition 7

$$E(X|Y = y) = \sum_i x_i \Pr(X = x_i|Y = y)$$

Example 9 *The expected value of milk yield (X) given genotype $Y = Aa$ is:*

$$\begin{aligned} E(X|Y = Aa) &= 100 \times \Pr(X = 100|Y = Aa) + 200 \times \Pr(X = 200|Y = Aa) \\ &\quad + 300 \times \Pr(X = 300|Y = Aa) \\ &= 100(0.2) + 200(0.6) + 300(0.2) \\ &= 200 \end{aligned}$$

2.9 Double Expectation Theorem

$$\underset{Y}{\mathrm{E}}[\mathrm{E}(X|Y)] = \mathrm{E}(X)$$

2.10 Proof of Double Expectation Theorem

From page 9, the conditional mean of X given $Y = y_j$ is

$$\mathrm{E}(X|Y = y_j) = \sum_i x_i \Pr(X = x_i|Y = y_j).$$

Note that $\mathrm{E}(X|Y = y_j)$ can be computed for every y_j in the sample space of Y. So, from the definition of expected value (page 7), the expected value of $\mathrm{E}(X|Y)$ is

$$\begin{aligned} \underset{Y}{\mathrm{E}}[\mathrm{E}(X|Y)] &= \sum_j \mathrm{E}(X|Y = y_j) \Pr(Y = y_j) \\ &= \sum_j [\sum_i x_i \Pr(X = x_i|Y = y_j)] \Pr(Y = y_j) \\ &= \sum_j \sum_i x_i \Pr(X = x_i, Y = y_j) \\ &= \sum_i x_i \sum_j \Pr(X = x_i, Y = y_j) \\ &= \sum_i x_i \Pr(X = x_i) \\ &= \mathrm{E}(X) \end{aligned}$$

Example 10

Genotype (Y)	$\mathrm{E}(X\|Y)$	$\Pr(Y)$
aa	140	0.25
Aa	200	0.5
AA	260	0.25

$$\begin{aligned}\underset{Y}{\mathrm{E}}[\mathrm{E}(X|Y)] &= \mathrm{E}(X|Y=aa)\times\Pr(Y=aa)+\mathrm{E}(X|Y=Aa)\times\Pr(Y=Aa)\\ &\quad+\mathrm{E}(X|Y=AA)\times\Pr(Y=AA)\\ &140(0.25)+200(0.5)+260(0.25)\\ &=200\end{aligned}$$

2.11 Useful Identity for Variance

Often, it is useful to write the variance as

$$\mathrm{Var}(X) = \underset{Y}{\mathrm{E}}[\mathrm{Var}(X|Y)] + \underset{Y}{\mathrm{Var}}[\mathrm{E}(X|Y)] \tag{1}$$

To prove the above identity, write the first term of (1) as

$$\begin{aligned}\underset{Y}{\mathrm{E}}[\mathrm{Var}(X|Y)] &= \underset{Y}{\mathrm{E}}\{\mathrm{E}(X^2|Y)-[\mathrm{E}(X|Y)]^2\}\\ &= \underset{Y}{\mathrm{E}}\{\mathrm{E}(X^2|Y)\}-\underset{Y}{\mathrm{E}}\{[\mathrm{E}(X|Y)]^2\}\\ &= \mathrm{E}(X^2)-\underset{Y}{\mathrm{E}}\{[\mathrm{E}(X|Y)]^2\}\end{aligned} \tag{2}$$

and second term of (1) as

$$\begin{aligned}\underset{Y}{\mathrm{Var}}[\mathrm{E}(X|Y)] &= \underset{Y}{\mathrm{E}}\{[\mathrm{E}(X|Y)]^2\}-\{\underset{Y}{\mathrm{E}}[\mathrm{E}(X|Y)]\}^2\\ &= \underset{Y}{\mathrm{E}}\{[\mathrm{E}(X|Y)]^2\}-[\mathrm{E}(X)]^2\end{aligned} \tag{3}$$

The sum of (2) and (3) gives $\mathrm{E}(X^2)-[\mathrm{E}(X)]^2$, which is the variance of X.

2.12 Statistical Independence

If random variables X and Y are independent,

$$\begin{aligned}\Pr(X=x|Y=y) &= \frac{\Pr(X=x,Y=y)}{\Pr(Y=y)}\\ &= Pr(X=x)\end{aligned}$$

Then it follows that

$$Pr(X = x, Y = y) = Pr(X = x) \Pr(Y = y)$$

2.13 Covariance

Definition 8

$$\begin{aligned} Cov(X, Y) &= E\{[X - E(X)][Y - E(Y)]\} \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

where

$$E(XY) = \sum_i \sum_j x_i y_j \Pr(X = x_i, Y = y_j)$$

2.13.1 Covariance Example

Example 11

Genotype (T)	Genotypic Value (G)	Phenotypic Value (P)	Probability
aa	140	100	0.175
aa	140	200	0.05
aa	140	300	0.025
Aa	200	100	0.1
Aa	200	200	0.3
Aa	200	300	0.1
AA	260	100	0.025
AA	260	200	0.05
AA	260	300	0.175

2.13.2 Computing Cov(G, P)

$$\begin{aligned} \mathrm{E}(GP) &= 140 \times 100 \times 0.175 + 140 \times 200 \times 0.05 + \cdots \\ &\quad + 260 \times 300 \times 0.175 \\ &= 41800 \end{aligned}$$

$$\begin{aligned} \mathrm{E}(G) &= 140 \times 0.175 + 140 \times 0.05 + \cdots \\ &\quad + 260 \times 0.175 \\ &= 200 \end{aligned}$$

$$\begin{aligned}\mathrm{E}(P) &= 100 \times 0.175 + 200 \times 0.05 + \cdots \\ &\quad + 300 \times 0.175 \\ &= 200\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(GP) &= \mathrm{E}(GP) - \mathrm{E}(G)\mathrm{E}(P) \\ &= 41800 - 200 \times 200 \\ &= 1800\end{aligned}$$

2.14 Covariance— Special Cases

$$\begin{aligned}\mathrm{Cov}(X, X) &= \mathrm{E}(XX) - \mathrm{E}(X)\mathrm{E}(X) \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)\mathrm{E}(X) \\ &= \mathrm{Var}(X)\end{aligned}$$

If X and Y are independent,

$$\begin{aligned}\mathrm{E}(X, Y) &= \sum_i \sum_j x_i y_j \Pr(X = x_i, Y = y_j) \\ &= \sum_i \sum_j x_i y_j \Pr(X = x_i) \Pr(Y = y_j) \\ &= [\sum_i x_i \Pr(X = x_i)][\sum_j y_j \Pr(Y = y_j)] \\ &= \mathrm{E}(X)\mathrm{E}(Y)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \\ &= \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(X)\mathrm{E}(Y) \\ &= 0\end{aligned}$$

2.15 Properties of Random Variables

For constants a and c and random variables X, Y, and Z:

$$\mathrm{E}(a) = a$$

$$\mathrm{E}(aX) = a\mathrm{E}(X)$$

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

$$\mathrm{E}(a + cX) = \mathrm{E}(a) + c\mathrm{E}(X)$$

$$\mathrm{Var}(a) = 0$$

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$$

$$\begin{aligned}\mathrm{Var}(a + X) &= \mathrm{Var}(a) + \mathrm{Var}(X) + 2\mathrm{Cov}(a, X) \\ &= \mathrm{Var}(X)\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X + Y + Z) = &\mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z) \\ &+ 2\mathrm{Cov}(X, Y) + 2\mathrm{Cov}(X, Z) + 2\mathrm{Cov}(Y, Z)\end{aligned}$$

2.16 Regression

Definition 9 *The regression of Y on X is:*

$$\hat{Y} = E(Y|X)$$

This is also called the best predictor of Y given X.

Regression model for Y:

$$Y = \hat{Y} + e$$

where

$$e = Y - \hat{Y}$$

is called the residual

2.16.1 Regression— Property 1

From the double expectation theorem,

$$\begin{aligned}\mathrm{E}(\hat{Y}) &= \underset{X}{\mathrm{E}}[\mathrm{E}(Y|X)] \\ &= \mathrm{E}(Y)\end{aligned}$$

The genotypic value (G) is the conditional expectation of the phenotypic value (P), given the genotype (T). So,

$$\begin{aligned}\mathrm{E}(G) &= \underset{T}{\mathrm{E}}[\mathrm{E}(P|T)] \\ &= \mathrm{E}(P)\end{aligned}$$

2.16.2 Regression— Property 2

The residual (e) has null expectation:

$$\begin{aligned}\mathrm{E}(e) &= \mathrm{E}(Y - \hat{Y}) \\ &= \mathrm{E}(Y) - \mathrm{E}(\hat{Y}) \\ &= \mathrm{E}(Y) - \mathrm{E}(Y) \\ &= 0\end{aligned}$$

2.16.3 Regression— Property 3

Can show that $\hat{Y}$ and e have null covariance. Because $\mathrm{E}(e) = 0$,

$$\begin{aligned}\mathrm{Cov}(\hat{Y}, e) &= \mathrm{E}(\hat{Y}e) \\ &= \underset{X}{\mathrm{E}}[\mathrm{E}(\hat{Y}e|X)] \\ &= \underset{X}{\mathrm{E}}[\hat{Y}\mathrm{E}(e|X)] \\ &= \underset{X}{\mathrm{E}}\{\hat{Y}[\mathrm{E}(Y|X) - \mathrm{E}(\hat{Y}|X)]\} \\ &= \underset{X}{\mathrm{E}}\{\hat{Y}[\hat{Y} - \hat{Y}]\} \\ &= \underset{X}{\mathrm{E}}[\hat{Y}(0)] \\ &= 0\end{aligned}$$

2.17 Regression Example

Example 12

$$P = G + E$$

where $G = \mathrm{E}(P|T)$

T	G	$E = (P - G)$	$\Pr(T)$	$\Pr(P\|T)$	$\Pr(T, P)$
aa	140	(100 - 140)	0.25	0.7	0.175
aa	140	(200 - 140)	0.25	0.2	0.05
aa	140	(300 - 140)	0.25	0.1	0.025
Aa	200	(100 - 200)	0.5	0.2	0.1
Aa	200	(200 - 200)	0.5	0.6	0.3
Aa	200	(300 - 200)	0.5	0.2	0.1
AA	260	(100 - 260)	0.25	0.1	0.025
AA	260	(200 - 260)	0.25	0.2	0.05
AA	260	(300 - 260)	0.25	0.7	0.175

$$
\begin{aligned}
E(GE) &= 140(100-140)(0.25)(0.7) \\
&\quad + 140(200-140)(0.25)(0.2) \\
&\quad + 140(300-140)(0.25)(0.1) \\
&\quad \vdots \\
&= (0.25)140\{[100(0.7)+200(0.2)+300(0.1)] \\
&\qquad - [140(0.7+0.2+0.1)]\} \\
&\quad \vdots \\
&= (0.25)140(140-140) \\
&\quad + (0.50)200(200-200) \\
&\quad + (0.25)160(160-160) \\
&= 0
\end{aligned}
$$

2.18 Correlation

Definition 10

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}$$

$$-1 \leq \mathrm{Cor}(X, Y) \leq 1$$

3 Single-Locus Inheritance

Most traits of economic importance are determined by a large number of loci. Before we study the inheritance of such traits, we will examine the inheritance at a single locus.The genetic constitution of a population for a single locus is completely described by the genotypic frequencies at that locus. However, genotypes are not directly transmitted from parents to offspring; rather, it is the genes that are transmitted. Therefore, it is useful to look at the the relationship between genotype frequencies and gene frequencies.

3.1 Genotype and Gene Frequencies

Consider a locus with two alleles A_1 and A_2. Let N_{ij} be the frequency of individuals with genotype A_iA_j. Then, the relative frequencies of the genotypes are

$$P_{11} = \frac{N_{11}}{N},$$

$$P_{12} = \frac{N_{12}}{N},$$

and

$$P_{22} = \frac{N_{22}}{N},$$

where $N = N_{11} + N_{12} + N_{22}$ is the total number of individuals. The relative frequencies of A_1 and A_2 are

$$\begin{aligned} p_1 &= \frac{2N_{11} + N_{12}}{2N} \\ &= P_{11} + \frac{1}{2}P_{12} \end{aligned} \tag{4}$$

and

$$\begin{aligned} p_2 &= \frac{2N_{22} + N_{12}}{2N} \\ &= P_{22} + \frac{1}{2}P_{12} \end{aligned} \tag{5}$$

Suppose that all individuals are equally likely to produce gametes (no selection) and that there is no mutation or migration. Then,

if a sufficiently large number of offspring are produced, the gene frequency in the offspring would be the same as the gene frequency in the parents. Further, if parents are sampled independently (this is often called random mating), then the genotypic frequencies are given by the Hardy-Weinberg Law.

3.2 Hardy-Weinberg Law

If:

1. mating is at random in large population
2. no selection, mutation, or migration

Then:

1. frequencies of genes and genotypes stay constant from generation to generation
2. simple relationship between gene frequencies in parents and genotype frequencies in offspring: if frequencies for two alleles A_1 and A_2 in parents are p_1 and p_2, then the frequencies for genotypes A_1A_1, A_1A_2 and A_2A_2 in the progeny are p_1^2, $2p_1p_2$, and p_2^2.

Thus, regardless of the genotypic frequencies in the parents, if a large number of progeny are produced, and there is no selection, mutation, or migration, the frequencies for genotypes A_1A_1, A_1A_2 and A_2A_2 in the progeny are p_1^2, $2p_1p_2$, and p_2^2. As we will see below, genotype and gamete frequencies for two loci do not reach equilibrium frequencies in one generation.

3.3 Two-Locus Gamete Frequencies

Consider locus A with alleles $A_1, A_2, A_3, \ldots$ and locus B with alleles $B_1, B_2, B_3, \ldots$. Let p_i^A be the frequency of allele A_i and p_j^B the frequency of B_j. In generation t, the probability that an individual x produces a gamete $g_x = A_xB_x$ with alleles $A_x = A_i$ and $B_x = B_j$ is denoted by p_{ij}^t. The gamete that x received from its mother is denoted A_mB_m and that it received from its father is denoted A_fB_f. We will now derive an expression for the gamete frequency in generation t in terms of the gamete frequency in generation $t-1$, the gene frequencies, and the recombination rate r between the two loci. The

gamete g_x can get alleles A_i and B_j in one of four mutually exclusive ways:

1. g_x is the maternal gamete $A_m B_m$ of x and $A_m = A_i, B_m = B_j$,
2. g_x is the paternal gamete $A_f B_f$ of x and $A_f = A_i, B_f = B_j$,
3. g_x is the recombinant $A_m B_f$ and $A_m = A_i, B_f = B_j$, or
4. g_x is the recombinant $A_f B_m$ and $A_f = A_i, B_m = B_j$

The probability for the first of these four events can be written as

$$\Pr(g_x = A_m B_m, A_m = A_i, B_m = B_j) = \Pr(g_x = A_m B_m) \Pr(A_m = A_i, B_m = B_j)$$

because we assume that $\Pr(g_x = A_m B_m)$ does not depend on the maternal haplotype. For example,

$$\Pr(g_x = A_m B_m) = 1/2(1 - r)$$

for maternal haplotype $(A_m = A_1, B_m = B_1)$ or $(A_m = A_1, B_m = B_2)$ or any other maternal haplotype. Substituting $\Pr(g_x = A_m B_m) = 1/2(1 - r)$ and $\Pr(A_m = A_i, B_m = B_j) = p_{ij}^{t-1}$ in the above gives

$$\Pr(g_x = A_m B_m, A_m = A_i, B_m = B_j) = 1/2(1 - r) p_{ij}^{t-1}.$$

Similarly,

$$\Pr(g_x = A_f B_f, A_f = A_i, B_f = B_j) = 1/2(1 - r) p_{ij}^{t-1},$$

$$\Pr(g_x = A_m B_f, A_m = A_i, B_f = B_j) = 1/2 r p_i^A p_j^B,$$

and

$$\Pr(g_x = A_f B_m, A_f = A_i, B_m = B_j) = 1/2 r p_i^A p_j^B.$$

Finally, the sum of these four probabilities gives

$$p_{ij}^t = \Pr(g_x = A_i B_j) = (1 - r) p_{ij}^{t-1} + r p_i^A p_j^B$$

3.4 Gametic Disequilibrium

$$\begin{aligned}
\Delta^t &= p_{ij}^t - p_i^A p_j^B \\
&= (1-r)p_{ij}^{t-1} + rp_i^A p_j^B - p_i^A p_j^B \\
&= (1-r)p_{ij}^{t-1} - (p_i^A p_j^B - rp_i^A p_j^B) \\
&= (1-r)p_{ij}^{t-1} - (1-r)p_i^A p_j^B \\
&= (1-r)(p_{ij}^{t-1} - p_i^A p_j^B) \\
&= (1-r)\Delta^{t-1} \\
&= (1-r)^2\Delta^{t-2} \\
&\vdots \\
&= (1-r)^t\Delta^0
\end{aligned}$$

Example 13 *Suppose* $r = 0.5$*, then*

$$\begin{aligned}
\Delta^{10} &= (1-r)^{10}\Delta^0 \\
&= \frac{1}{1024}\Delta^0
\end{aligned}$$

Example 14 *Suppose* $r = 0.1$*, then*

$$\begin{aligned}
\Delta^{10} &= (1-r)^{10}\Delta^0 \\
&= \frac{9^{10}}{10^{10}}\Delta^0 \\
&= 0.349\Delta^0
\end{aligned}$$

3.5 Change in Gene Frequency Due to Migration

Consider a large population with a proportion m of immigrants. Let frequency of A_2 be q_0 in the natives and q_m in the immigrants. Then the frequency q_1 of A_2 in the mixed population is

$$\begin{aligned}
q_1 &= mq_m + (1-m)q_0 \\
&= m(q_m - q_0) + q_0
\end{aligned}$$

3.6 Change in Gene Frequency Due to Mutation

Consider a locus where A_1 is the normal allele and A_2 is the mutant. Suppose A_1 mutates to A_2 with probability u and A_2 mutates to A_1 with probability v. In generation 0, the frequency of A_1 is denoted p_0 and the frequency of A_2 is denoted q_0. Then, in the absence of migration, selection and drift, the frequency of A_2 in generation 1 is

$$q_1 = (1 - v)q_0 + up_0,$$

and the change in frequency of A_2 is

$$\Delta_q = up_0 - vq_0$$

At equilibrium, the probability of A_1 mutating to A_2 will be equal to the probability of A_2 mutating to A_1. Thus, for the equilibrium frequency p of A_1 and q of A_2,

$$pu = qv.$$

Substituting $(1 - q)$ for p gives

$$(1 - q)u = qv,$$

and solving for q gives

$$q = \frac{u}{u + v}$$

Mutation rate v from the mutant to the normal has been observed to be much lower than the rate u from the normal to the mutant. Suppose that

$$v = \frac{u}{10}.$$

Then, the equilibrium frequency of A_2 is

$$\begin{aligned} q &= \frac{u}{u + v} \\ &= \frac{u}{u + \frac{u}{10}} \\ &= \frac{10}{11}. \end{aligned}$$

However, mutant alleles are very rare. As we will see later, this is due to selection.

3.7 Change in Gene Frequency Due to Selection

We will consider a locus with two alleles A_1 and A_2, and assume there is no migration, mutation or drift. In generation 0, the allele frequencies at conception are

$$p = \Pr(A_1)$$

and

$$q = \Pr(A_2)$$

Suppose N zygotes are produced by random mating. The genotypic numbers at conception are

$$N_{11} = Np^2,$$

$$N_{12} = N2pq,$$

and

$$N_{22} = Nq^2$$

Now we will allow these zygotes to have different levels of fertility.

Definition 11 *Fitness W_{ij} of genotype A_iA_j is the average number of gametes transmitted to the next generation by zygotes with genotype A_iA_j.*

So, the average number of gametes transmitted to the next generation is

$$Np^2W_{11}$$

for zygotes of genotype A_1A_1,

$$N2pqW_{12}$$

for zygotes of genotype A_1A_2, and

$$Nq^2W_{22}$$

for zygotes for genotype A_2A_2.

The frequency of allele A_1 among these gametes, which are transmitted to generation 1, is

$$\begin{aligned} p_1 &= \frac{2Np^2W_{11} + N2pqW_{12}}{2Np^2W_{11} + 2N2pqW_{12} + 2Nq^2W_{22}} \\ &= \frac{p^2W_{11} + pqW_{12}}{p^2W_{11} + 2pqW_{12} + q^2W_{22}} \\ &= \frac{p^2W_{11} + pqW_{12}}{\bar{W}^*} \\ &= \frac{p^2 + pq\frac{W_{12}}{W_{11}}}{\bar{W}}, \end{aligned} \tag{6}$$

where

$$\bar{W}^* = p^2W_{11} + 2pqW_{12} + q^2W_{22}$$

is the average fitness, $\frac{W_{12}}{W_{11}}$ is the fitness of genotype A_1A_2 relative to the fitness of A_1A_1, and $\bar{W} = \frac{\bar{W}^*}{W_{11}}$ is the average fitness relative to the fitness of A_1A_1. Similarly, the frequency of allele A_2 is

$$q_1 = \frac{q^2\frac{W_{22}}{W_{11}} + pq\frac{W_{12}}{W_{11}}}{\bar{W}}, \tag{7}$$

where $\frac{W_{22}}{W_{11}}$ is the relative fitness of genotype A_2A_2.
The relative fitness for A_1A_2 can be expressed as

$$\frac{W_{12}}{W_{11}} = 1 - hs,$$

and for A_2A_2 as

$$\frac{W_{22}}{W_{11}} = 1 - s,$$

where hs is the coefficient of selection for A_1A_2 and s is the coefficient of selection for A_2A_2. Now, equation (6) can be written

as

$$\begin{aligned} p_1 &= \frac{p^2 + pq(1-hs)}{\bar{W}} \\ &= \frac{p^2 + pq - pqhs}{\bar{W}} \\ &= \frac{p(p+q-qhs)}{\bar{W}} \\ &= \frac{p(1-qhs)}{\bar{W}}. \end{aligned} \tag{8}$$

Similarly, the frequency of allele A_2 in generation 1 zygotes is

$$q_1 = \frac{q - sq^2 - hspq}{\bar{W}} \tag{9}$$

The change in frequency for allele A_1 is

$$\begin{aligned} \Delta_p &= p_1 - p \\ &= \frac{p(1-qhs)}{\bar{W}} - p \\ &= \frac{p[(1-qhs) - \bar{W}]}{\bar{W}}. \end{aligned} \tag{10}$$

Note that $\bar{W}$ can be written as

$$\begin{aligned} \bar{W} &= p^2 + 2pq(1-hs) + q^2(1-s) \\ &= 1 - 2hspq - sq^2. \end{aligned} \tag{11}$$

Substituting (11) in the numerator of (10) and rearranging gives

$$\Delta_p = \frac{pqs}{\bar{W}}[q + h(p-q)]. \tag{12}$$

Because $p + q = 1$, the change in frequency for allele A_2 is

$$\begin{aligned} \Delta_q &= -\Delta_p \\ &= -\frac{pqs}{\bar{W}}[q + h(p-q)]. \end{aligned} \tag{13}$$

For the overdominant case, the relative fitness of each homozygote is written as

$$\frac{W_{11}}{W_{12}} = 1 - s_1$$

and

$$\frac{W_{22}}{W_{12}} = 1 - s_2.$$

Then, the frequency of A_1 is

$$p_1 = \frac{p - s_1 p^2}{1 - s_1 p^2 - s_2 q^2}, \tag{14}$$

and the frequency of A_2 is

$$q_1 = \frac{q - s_2 q^2}{1 - s_1 p^2 - s_2 q^2}. \tag{15}$$

The change in frequency for the A_2 allele is

$$\begin{aligned} \Delta_q &= q_1 - q \\ &= \frac{pq(s_1 p - s_2 q)}{1 - s_1 p^2 - s_2 q^2} \end{aligned} \tag{16}$$

The above formulae can be used to examine the effectiveness of selection for different modes of inheritance. For example, if allele A_2 is a dominant lethal, the frequency of A_2 in the next generation can be computed from (9) by putting $s = 1$ and $h = 1$, which gives

$$\begin{aligned} q_1 &= \frac{q - q^2 - pq}{\bar{W}} \\ &= \frac{q - q(q + p)}{\bar{W}} \\ &= \frac{q - q}{\bar{W}} \\ &= 0. \end{aligned} \tag{17}$$

This is because none of the A_2 alleles is transmitted to the next generation. However, if A_2 is a recessive lethal, putting $s = 1$ and $h = 0$ in (9) gives

$$\begin{aligned} q_1 &= \frac{q - q^2}{\bar{W}} \\ &= \frac{q(1 - q)}{1 - q^2} \\ &= \frac{q}{1 + q} \\ &\geq 0. \end{aligned} \tag{18}$$

This is because all the A_2 alleles in the heterozygotes are transmitted to the next generation. Plotting response to selection as a function of gene frequency shows that

1. response to selection is greatest when gene frequencies are intermediate, and
2. response to selection for a rare recessive lethal allele is very low.

The number of generations required to change the frequency of a rare recessive lethal by a specified amount can be computed as follows. Using (18), the frequency in generation 2 can be written as

$$\begin{aligned} q_2 &= \frac{q_1}{1+q_1} \\ &= \frac{\frac{q}{1+q}}{1+\frac{q}{1+q}} \\ &= \frac{\frac{q}{1+q}}{\frac{1+q+q}{1+q}} \\ &= \frac{q}{1+2q}, \end{aligned} \tag{19}$$

and the frequency in generation 3 can be written as

$$\begin{aligned} q_3 &= \frac{q_2}{1+q_2} \\ &= \frac{q_1}{1+2q_1} \\ &= \frac{\frac{q}{1+q}}{1+\frac{2q}{1+q}} \\ &= \frac{\frac{q}{1+q}}{\frac{1+q+2q}{1+q}} \\ &= \frac{q}{1+3q}. \end{aligned} \tag{20}$$

Continuing this process, the frequency at generation t is

$$q_t = \frac{q}{1+tq} \tag{21}$$

Now, solving for t from (21) gives

$$\begin{aligned} 1 + tq &= \frac{q}{q_t} \\ t &= \frac{1}{q_t} - \frac{1}{q} \end{aligned} \tag{22}$$

Example 2.2 from Falconer and Mackay Suppose Albinism is due to a single recessive locus. The present frequency of Albinism is

$$\Pr(A_2A_2) = \frac{1}{20,000}$$

Assuming Hardy-Weinberg equilibrium, the frequency q of A_2 is

$$\begin{aligned} q &= \sqrt{\frac{1}{20,000}} \\ &= \frac{1}{141} \end{aligned}$$

To reduce the frequency of Albinism to $\frac{1}{40,000}$, the frequency of A_2 must be reduced to q_t

$$\begin{aligned} q_t &= \sqrt{\frac{1}{40,000}} \\ &= \frac{1}{200} \end{aligned}$$

Suppose this is to be achieved by preventing Albino's to reproduce. Then, from (22), the number of generations required to achieve this is

$$\begin{aligned} t &= \frac{1}{q_t} - \frac{1}{q} \\ &= 200 - 141 \\ &= 59 \end{aligned}$$

At 25 years per generation this would take 1475 years.

3.8 Equilibrium Between Mutation and Selection

Suppose A_2 is the mutant allele. Then, selection against A_2 will tend to reduce its frequency. But, mutation from A_1 to A_2 (at rate u) will

keep it from being completely lost. Since mutations are rare, we can expect the frequency of A_2 to be low. So, reverse mutations from A_2 to A_1 will be very rare and will be ignored in the following. The frequency of A_1 in the zygotes of the present generation is denoted by p. Then, from (8), in the absence of mutation, the frequency of A_1 in the zygotes of the next generation is

$$\frac{p(1-qhs)}{1-2pqhs-sq^2}$$

However, between generations, a fraction u of the A_1 alleles will mutate to A_2. So, with mutation, the frequency p_1 of A_1 in the zygotes of the next generations is

$$p_1 = \frac{p(1-qhs)}{1-2pqhs-sq^2}(1-u) \tag{23}$$

At equilibrium, $p_1 = p$. So, we get

$$\begin{aligned} p &= \frac{p(1-qhs)}{1-2pqhs-sq^2}(1-u) \\ 1-2(1-q)qhs-sq^2 &= (1-qhs)(1-u) \\ 1-2qhs+2q^2hs-sq^2 &= 1-u-qhs(1-u) \\ s(2h-1)q^2-hs(1+u)q+u &= 0 \end{aligned} \tag{24}$$

Case: A_1 is dominant Then, $h = 0$ and (24) reduces to

$$\begin{aligned} -sq^2+u &= 0 \\ u &= sq^2. \end{aligned} \tag{25}$$

Solving for q from (25) gives the equilibrium frequency for A_2

$$q = \sqrt{\frac{u}{s}} \tag{26}$$

Case: no dominance Here, $h = \frac{1}{2}$ and (24) becomes

$$\begin{aligned} -\frac{s(1+u)q}{2}+u &= 0 \\ q &= \frac{2u}{s(1+u)} \\ &\approx \frac{2u}{s} \end{aligned} \tag{27}$$

Case: A_2 dominant Then $h = 1$, and because $(1 + u) \approx 1$, (24) reduces to

$$\begin{aligned} sq^2 - sq + u &= 0 \\ sq(q-1) &= -u \\ sq(1-q) &= u \\ spq &= u \\ pq &= \frac{u}{s} \end{aligned} \tag{28}$$

Under Hardy-Weinberg frequencies, the mutant phenotype has frequency

$$H = 2pq + q^2$$

If q is very small, $H \approx 2pq$. Thus, the frequency of the mutant phenotype is approximately $\frac{2u}{s}$.

Estimation of mutation rate— Example 2.3 from F&M Dominant dwarfism is a dominant abnormality. So, A_2 is dominant and $h = 1$. The frequency of this type of dwarfism in Denmark is 10.7×10^{-5}. Their relative fitness is

$$\begin{aligned} (1-s) &= \frac{\text{average number of children from dwarfs}}{\text{average number of children from normal sibs}} \\ &= 0.196, \end{aligned}$$

and so

$$\begin{aligned} s &= 1 - 0.196 \\ &= 0.804 \end{aligned}$$

Now, if we use the frequency of dwarfs for H in (28), the mutation rate can be estimated as

$$\begin{aligned} u &= \frac{sH}{2} \\ &= \frac{0.804 \times 10.7 \times 10^{-5}}{2} \\ &= 4.3 \times 10^{-5} \end{aligned}$$

Equilibrium frequencies Given that mutation rates are about 10^{-5}, only mild selection is needed to keep the frequency of the mutant from very low. For example, suppose A_2 is recessive,

$$u = 10^{-5},$$

and

$$s = 0.1.$$

Then, from (26), the equilibrium frequency is

$$\begin{aligned} q &= \sqrt{\frac{u}{s}} \\ &= \sqrt{\frac{10^{-5}}{0.1}} \\ &= \frac{1}{100} \end{aligned}$$

3.9 Equilibrium Under Overdominance

At equilibrium, $\Delta_q = 0$. Thus, from (16),

$$\begin{aligned} s_1 p &= s_2 q \\ s_1(1-q) &= s_2 q \\ q(s_1 + s_2) &= s_1 \\ q &= \frac{s_1}{s_1 + s_2} \end{aligned} \tag{29}$$

Note that the equilibrium frequency does not depend on the degree of overdominance. It depends on the fitness of one homozygote relative to the other.

3.10 Change in Gene Frequency Due to Drift

When a finite number of gametes is sampled from the parental population, the gene frequency in the gametes will "randomly" deviate from the frequency in the parents. This process is referred to as genetic drift. We will now examine the consequences of drift and the relationship between "sample size" and the magnitude of drift. An ideal population with simplified structure is employed to model the process of drift. In the ideal population, we assume:

1. no mutation, migration, or selection,
2. generations do not overlap
3. population size N is constant across generations
4. mating is at random including self-fertilization

Further, in the following we will assume all loci have two alleles.

3.10.1 Mean and Variance of Gene Frequency

Consider all loci that have frequency q_0 for allele 2 in generation 0. Suppose $2N$ gametes are randomly sampled from generation 0 to produce N individuals in generation 1. At any one of these loci, let Y be the number of "2" alleles in generation 1. Because each allele can be "2" with probability q_0 and because the $2N$ gametes are sampled independently, $Y \sim \text{Binomial}(2N, q_0)$ (see example 19). The frequency q_1 of allele 2 in generation 1 is

$$q_1 = \frac{Y}{2N}.$$

The expected value of q_1 is

$$\begin{aligned} \mathrm{E}(q_1) &= \frac{\mathrm{E}(Y)}{2N} \\ &= \frac{2Nq_0}{2N} \\ &= q_0, \end{aligned}$$

and the variance of q_1 is

$$\begin{aligned} \mathrm{Var}(q_1) &= \frac{\mathrm{Var}(Y)}{(2N)^2} \\ &= \frac{2Nq_0(1-q_0)}{(2N)^2} \\ &= \frac{q_0(1-q_0)}{2N}. \end{aligned} \tag{30}$$

So, among all loci that had frequency q_0 for allele 2 in generation 0, some would have a higher frequency and others would have a lower frequency in generation 1. But, the distribution of frequencies

across loci in generation 1 would be centered at q_0. Further, from (30), the spread of this distribution would be inversely related to the population size N.
Now, $2N$ alleles are randomly sampled from generation 1 to produce N individuals in generation 2. Among all loci that had frequency q_1 for allele 2 in generation 1 parents, the frequency q_2 of allele 2 in generation 2 is distributed as

$$q_2 \sim \frac{\text{Binomial}(2N, q_1)}{2N}.$$

Among these loci that had frequency q_1 for allele 2 in generation 1, the expected value for q_2 is

$$\mathrm{E}(q_2|q_1) = q_1,$$

and the variance for q_2 is

$$\mathrm{Var}(q_2|q_1) = \frac{q_1(1-q_1)}{2N}.$$

Now, the expected value of q_2 among all loci that had frequency q_0 for allele 2 in generation 0 is given by

$$\begin{aligned}\mathrm{E}(q_2) &= \mathop{\mathrm{E}}_{q_1}[\mathrm{E}(q_2|q_1)] \\ &= \mathrm{E}(q_1) \\ &= q_0,\end{aligned}$$

and the variance of q_2 among these loci is

$$\begin{aligned}\mathrm{Var}(q_2) &= \mathop{\mathrm{E}}_{q_1}\mathrm{Var}(q_2|q_1) + \mathop{\mathrm{Var}}_{q_1}\mathrm{E}(q_2|q_1) \\ &= \mathrm{E}[\frac{q_1(1-q_1)}{2N}] + \mathrm{Var}(q_1) \\ &= \frac{1}{2N}\mathrm{E}(q_1 - q_1^2) + \mathrm{Var}(q_1) \\ &= \frac{1}{2N}[q_0 - q_0^2 - \mathrm{Var}(q_1)] + \mathrm{Var}(q_1) \\ &= \mathrm{Var}(q_1) + \mathrm{Var}(q_1)(1 - \frac{1}{2N}) \\ &= \mathrm{Var}(q_1)[1 + (1 - \frac{1}{2N})]\end{aligned}$$

Similarly, among all loci that had frequency q_2 for allele 2 in generation 2, the frequency q_3 of allele 2 in generation 3 is distributed as

$$q_3 \sim \frac{\text{Binomial}(2N, q_2)}{2N}.$$

Among these loci that had frequency q_2 for allele 2 in generation 2, the expected value for q_3 is

$$\text{E}(q_3|q_2) = q_2,$$

and the variance for q_3 is

$$\text{Var}(q_3|q_2) = \frac{q_2(1-q_2)}{2N}.$$

The expected value of q_3 among all loci that had frequency q_0 for allele 2 in generation 0 is given by

$$\begin{aligned}
\text{E}(q_3) &= \underset{q_2}{\text{E}}[\text{E}(q_3|q_2)] \\
&= \text{E}(q_2) \\
&= q_0,
\end{aligned}$$

and the variance of q_3 among these loci is

$$\begin{aligned}
\text{Var}(q_3) &= \underset{q_2}{\text{E}}\text{Var}(q_3|q_2) + \underset{q_2}{\text{Var}}\text{E}(q_3|q_2) \\
&= \text{E}[\frac{q_2(1-q_2)}{2N}] + \text{Var}(q_2) \\
&= \frac{1}{2N}\text{E}(q_2 - q_2^2) + \text{Var}(q_2) \\
&= \frac{1}{2N}[q_0 - q_0^2 - \text{Var}(q_2)] + \text{Var}(q_2) \\
&= \text{Var}(q_1) + \text{Var}(q_2)(1 - \frac{1}{2N}) \\
&= \text{Var}(q_1) + \text{Var}(q_1)[1 + (1 - \frac{1}{2N})](1 - \frac{1}{2N}) \\
&= \text{Var}(q_1)[1 + (1 - \frac{1}{2N}) + (1 - \frac{1}{2N})^2]
\end{aligned}$$

Similarly, the expected value of q_4 in generation 4 among all loci that had frequency q_0 for allele 2 in generation 0 is

$$\text{E}(q_4) = q_0$$

and variance of q_4 is

$$\begin{aligned}\mathrm{Var}(q_4) &= \mathrm{Var}(q_1) + \mathrm{Var}(q_3)(1 - \frac{1}{2N}) \\ &= \mathrm{Var}(q_1) + \mathrm{Var}(q_1)[1 + (1 - \frac{1}{2N}) + (1 - \frac{1}{2N})^2](1 - \frac{1}{2N}) \\ &= \mathrm{Var}(q_1)[1 + (1 - \frac{1}{2N}) + (1 - \frac{1}{2N})^2 + (1 - \frac{1}{2N})^3]\end{aligned}$$

In generation t, the expected value of q_t is

$$\mathrm{E}(q_t) = q_0$$

and variance of q_t is

$$\begin{aligned}\mathrm{Var}(q_t) &= \mathrm{Var}(q_1)[1 + (1 - \frac{1}{2N}) + (1 - \frac{1}{2N})^2 + \cdots + (1 - \frac{1}{2N})^{t-1}] \\ &= \frac{q_0(1-q_0)}{2N}[1 + (1 - \frac{1}{2N}) + (1 - \frac{1}{2N})^2 + \cdots + (1 - \frac{1}{2N})^{t-1}]\end{aligned}$$

Using (118) in the above equation gives

$$\begin{aligned}\mathrm{Var}(q_t) &= \frac{q_0(1-q_0)}{2N}[\frac{1 - (1 - \frac{1}{2N})^t}{1 - (1 - \frac{1}{2N})}] \\ &= q_0(1-q_0)[1 - (1 - \frac{1}{2N})^t]\end{aligned} \tag{31}$$

It can be shown that as t goes to infinity, q_t is either one or zero, i.e., q_t becomes a Bernoulli random variable. But, we also know that at any generation the expected value of q_t is q_0. The expected value of a Bernoulli random variable is equal to the probability that it is equal to one. Thus, the probability that $q_t = 1$, which is the probability of fixation of allele 2, is equal to q_0. So, among all loci that started out at a frequency of q_0 for allele 2, after a very large number of generations, a proportion q_0 of loci will have a frequency of 1 for allele 2 and a proportion $(1 - q)$ will have a frequency of 0 for this allele. Note that as t goes to infinity, the variance of q_t computed from (31) is $q_0(1-q_0)$, which is the variance of a Bernoulli random variable that is equal to one with probability q_0.

3.10.2 Distribution of Gene Frequency

Let Y_t be the number of "2" alleles in generation t. So,

$$Y_1 \sim \text{Binomial}(2N, q_0)$$

and

$$q_1 = \frac{Y_1}{2N}$$

So, for example,

$$\Pr(q_1) = \Pr(Y_1 = 2Nq_1).$$

Thus, the distribution of q_t is easily obtained from the distribution of Y_t.
Recall that

$$(Y_2|Y_1 = y_1) \sim \text{Binomial}(2N, q_1 = \frac{y_1}{2N}).$$

Thus, the joint distribution of Y_1 and Y_2 can be written as

$$Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1) \Pr(Y_2 = y_2|Y_1 = y_1),$$

and the marginal distribution of Y_2 is

$$\begin{aligned} \Pr(Y_2 = y_2) &= \sum_{y_1=0}^{2N} \Pr(Y_1 = y_1, Y_2 = y_2) \\ &= \sum_{y_1=0}^{2N} \Pr(Y_1 = y_1) \Pr(Y_2 = y_2|Y_1 = y_1) \end{aligned} \tag{32}$$

At generation t,

$$(Y_t|Y_{t-1} = y_{t-1}) \sim \text{Binomial}(2N, q_{t-1} = \frac{y_{t-1}}{2N}).$$

So, the joint distribution of Y_t and Y_{t-1} can be written as

$$Pr(Y_{t-1} = y_{t-1}, Y_t = y_t) = \Pr(Y_{t-1} = y_{t-1}) \Pr(Y_t = y_t|Y_{y-1} = y_{t-1}),$$

and the marginal distribution of Y_t is

$$\begin{aligned} \Pr(Y_t = y_t) &= \sum_{y_{t-1}=0}^{2N} \Pr(Y_{t-1} = y_{t-1}, Y_t = y_t) \\ &= \sum_{y_{t-1}=0}^{2N} \Pr(Y_{t-1} = y_{t-1}) \Pr(Y_t = y_t|Y_{t-1} = y_{t-1}) \end{aligned} \tag{33}$$

3.10.3 Mean of Genotype Frequencies

We have seen that the expected value of gene frequency q_t at generation t is q_0. Thus, expected gene frequency is constant across generations. As shown below, with random mating in a finite population, genotypic frequencies do not stay constant.
Under random mating, the frequency of the homozygous genotype for allele "2" at a random locus is given by q_t^2. The mean (expected value) of this frequency is

$$\begin{aligned} \mathrm{E}(q_t^2) &= [\mathrm{E}(q_t)]^2 + \mathrm{Var}(q_t) \\ &= q_0^2 + \mathrm{Var}(q_t) \end{aligned} \tag{34}$$

Similarly, the expected frequency of the homozygous genotype for allele "1" in generation t is

$$\begin{aligned} \mathrm{E}(p_t^2) &= [\mathrm{E}(p_t)]^2 + \mathrm{Var}(p_t) \\ &= p_0^2 + \mathrm{Var}(q_t) \end{aligned} \tag{35}$$

because $p_t = 1 - q_t$, and so $\mathrm{Var}(p_t) = \mathrm{Var}(q_t)$. The frequency of the heterozygous genotype at a random locus is $2(1-q_t)q_t$, and the expected value of this frequency is

$$\begin{aligned} \mathrm{E}[2(1-q_t)q_t] &= 2q_0 - 2\mathrm{E}(q_t^2) \\ &= 2q_0 - 2q_0^2 - 2\mathrm{Var}(q_t) \\ &= 2q_0(1-q_0) - 2\mathrm{Var}(q_t) \end{aligned} \tag{36}$$

From (31) we see that $\mathrm{Var}(q_t)$ increases each generation. Thus, from (34) and (35), with each generation of random mating, the expected frequency of homozygotes increases, and from (36), the expected frequency of heterzygotes decreases. As t goes to infinity, $\mathrm{Var}(q_t)$ becomes $q_0(1-q_0)$. Using this limiting value of the variance in (36) shows that in the limit each locus becomes homozygous for either the "1" or "2" allele. As shown below, these changes in genotype frequencies can be expressed in terms of inbreeding.

3.10.4 Inbreeding

Definition 12 *Mating individuals that are related results in inbreeding.*

Random mating in a finite population results in inbreeding.

Definition 13 *Two alleles at a locus are identical by descent (IBD) if they are both copies of a single ancestral allele. Alleles that are not IBD are said to be independent in descent.*

The coefficient of inbreeding denoted by F is the probability that the two alleles at a locus are identical by descent. In computing F, all founders are assumed to be unrelated and their alleles are assumed to be independent in descent.

3.10.5 Inbreeding in Ideal Population

Consider an ideal population of size N. It is assumed that all alleles in generation 0 are independent in descent. Each allele in generation 1 is a copy of one of the $2N$ alleles in generation 0. Thus, the probability that two randomly sampled alleles in generation 1 are both copies of the same allele from generation 0 is $\frac{1}{2N}$. This is the probability that two randomly sampled alleles in generation 1 are IBD, because all alleles in generation 0 are assumed to be independent in descent. Thus, the coefficient of inbreeding in generation 1 is

$$F_1 = \frac{1}{2N}$$

In generation 2, the probability that two randomly sampled alleles are both copies of the same allele of the previous generation is $\frac{1}{2N}$, and the probability that they are copies of different alleles of generation 1 is $(1 - \frac{1}{2N})$. However, two random alleles of generation 1 may be IBD with probability F_1. So, the coefficient of inbreeding in generation 2 is

$$F_2 = \frac{1}{2N} + (1 - \frac{1}{2N})F_1$$

In general, there are two ways in which two alleles in generation t can be identical by descent:

1. both are copies of the same allele of generation $t - 1$, or
2. they are copies of different alleles of generation $t - 1$ that are IBD.

The probability for the first of these is $\frac{1}{2N}$ and the probability for the second is $(1-\frac{1}{2N})F_{t-1}$. Thus, the inbreeding coefficient in generation

t is

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1} \tag{37}$$

As shown below, the coefficient of inbreeding can also be written as

$$F_t = 1 - (1 - \Delta_F)^t, \tag{38}$$

where

$$\begin{aligned} \Delta_F &= \frac{F_t - F_{t-1}}{1 - F_{t-1}} \\ &= \frac{1}{2N}. \end{aligned} \tag{39}$$

This is the change in F in generation t relative to the remaining possible change. Rearranging (39) gives

$$1 - \Delta_F = \frac{1 - F_t}{1 - F_{t-1}},$$

and so

$$\begin{aligned} (\frac{1 - F_1}{1 - F_0})(\frac{1 - F_2}{1 - F_1}) \cdots (\frac{1 - F_t}{1 - F_{t-1}}) &= (1 - \Delta_F)^t \\ \frac{1 - F_t}{1 - F_0} &= (1 - \Delta_F)^t \end{aligned}$$

Because all alleles in generation 0 are independent in descent, F_0 is null. Thus, the above euqation gives

$$\begin{aligned} F_t &= 1 - (1 - \Delta_F)^t \\ &= 1 - (1 - \frac{1}{2N})^t \end{aligned} \tag{40}$$

Now, from (31) and (38) the variance of gene frequencies can be written as

$$\begin{aligned} \mathrm{Var}(q_t) &= q_0(1 - q_0)[1 - (1 - \frac{1}{2N})^t] \\ &= q_0(1 - q_0)F_t \end{aligned} \tag{41}$$

The expected frequency of genotypes can be expressed in terms of F as follows. Let the maternal allele at locus A be denoted A^M and

the paternal allele A^P. These two alleles are IBD with probability F_t or independent in descent with probability $(1 - F_t)$. If they are independent in descent, the expected frequencies of A_1A_1, A_1A_2 and A_2A_2 are given by the Hardy-Weinberg principle as: $p_0^2, 2p_0q_0$, and q_0^2. If the maternal and paternal alleles are IBD, the probability of A_1A_1 can be written as

$$\begin{aligned}\Pr(A^M = A_1, A^P = A_1|IBD) &= \Pr(A^M = A_1)\Pr(A^P = A_1|A^M = A_1, IBD) \\ &= p_0\end{aligned} \tag{42}$$

So, the unconditional probability of A_1A_1 is

$$\Pr(A_1A_1) = p_0^2(1 - F_t) + p_0F_t.$$

Similarly, the probability of A_2A_2 is

$$\Pr(A_2A_2) = q_0^2(1 - F_t) + q_0F_t$$

Note that if the maternal and paternal alleles are IBD they cannot be heterozygous. Thus, the probability of the heterozugous genotype is

$$\Pr(A_1A_2) = 2p_0q_0(1 - F_t).$$

3.11 Drift Under Less Simplified Conditions

Consider a breeding population P that does not conform to the assumptions of the "ideal" population. Suppose we can compute the rate of inbreeding Δ_{F_P} for population P.

Definition 14 *The size N_e of an ideal population that has the same rate of inbreeding as population P is the effective population size for P.*

Thus, the changes in gene and genotype frequencies in P due to drift will be equivalent to these changes in an ideal population of size N_e. From (39), the effective population size for a population P is given by

$$N_e = \frac{1}{2\Delta_{F_P}}. \tag{43}$$

3.11.1 No Self-fertilization

We will now compute the rate of inbreeding for a random mating population that excludes self-fertilization. To do so, the concept of coancestry is used.

Definition 15 *The coefficient of coancestry between individuals X and Y is the probability that a randomly sampled allele from X is IBD to a randomly sampled allele from Y.*

Let g_t denote the coefficient of coancestry between two randomly sampled individuals of generation t. Then, under random mating, the inbreeding coefficient in generation t is

$$F_t = g_{t-1}. \tag{44}$$

Let Q_t denote the probability that two alleles sampled from different individuals in generation t originate in the same parent in generation $t-1$. Given that the alleles are from the same parent of generation $t-1$, the probability that they are IBD is $\frac{1+F_{t-1}}{2}$. The probability that two alleles sampled from different individuals originate in different parents of generation $t-1$ is $(1-Q_t)$. Given that the alleles are from different parents of generation $t-1$, the probability that they are IBD is g_{t-1}. Thus, the unconditional probability that two alleles sampled from different individuals are IBD is

$$g_t = Q_t\frac{(1+F_{t-1})}{2} + (1-Q_t)g_{t-1} \tag{45}$$

Using (44) in (45), the coefficient of inbreeding in generation t is written as

$$\begin{aligned} F_t &= Q_{t-1}\frac{(1+F_{t-2})}{2} + (1-Q_{t-1})F_{t-1} \\ &= F_{t-1} + (1-2F_{t-1}+F_{t-2})\frac{Q_{t-1}}{2}. \end{aligned} \tag{46}$$

Now, the rate of inbreeding can be written as

$$\begin{aligned} \Delta_{F_P} &= \frac{F_t - F_{t-1}}{1-F_{t-1}} \\ &= [\frac{1-F_{t-1}-(F_{t-1}-F_{t-2})}{1-F_{t-1}}]\frac{Q_{t-1}}{2}, \end{aligned} \tag{47}$$

and using the approximation

$$F_{t-1} - F_{t-2} \approx \Delta_{F_P}(1 - F_{t-1})$$

gives

$$\begin{aligned} \Delta_{F_P} &\approx [\frac{(1 - F_{t-1}) - \Delta_{F_P}(1 - F_{t-1})}{(1 - F_{t-1})}]\frac{Q_{t-1}}{2} \\ &\approx (1 - \Delta_{F_P})\frac{Q_{t-1}}{2} \end{aligned} \tag{48}$$

Rearranging (48) gives

$$\Delta_{F_P} \approx \frac{1}{\frac{2}{Q_{t-1}} + 1} \tag{49}$$

Now from (43), the effective population size when selfing is excluded becomes

$$\begin{aligned} N_e &= \frac{1}{2\Delta_{F_P}} \\ &\approx \frac{1}{Q_{t-1}} + \frac{1}{2} \end{aligned} \tag{50}$$

Under random mating,

$$Q_t = \frac{1}{N},$$

and thus the effective population size when selfing is excluded becomes

$$N_e \approx N + \frac{1}{2}$$

3.11.2 Unequal Numbers of Males and Females

Consider a population where N_m males are randomly mated to N_f females. The effective population size for such a population can be computed using (50). However, because $N_m \neq N_f$, $Q_t \neq \frac{1}{N_m+N_f}$. Recall that Q_t is the probability that two alleles, say x and y, sampled from different individuals in generation t originate in the same parent in generation $t-1$. Note that even though $N_m \neq N_f$, half the alleles in generation t originate from males in generation $t-1$.

Therefore, the probability that x and y are both from males of the previous generation is $\frac{1}{4}$. Now, given that x and y are both from males, the probability that they are from the same individual is $\frac{1}{N_m}$. Thus, the unconditional probability that x and y are both from the same male is $\frac{1}{4N_m}$. Similarly, the probability that x and y are from the same female is $\frac{1}{4N_f}$. So, the probability that x and y are from the same parent is

$$Q_t = \frac{1}{4N_m} + \frac{1}{4N_f} \tag{51}$$

Substituting (51) in (50) gives

$$N_e = \frac{1}{\frac{1}{4N_m} + \frac{1}{4N_f}} + \frac{1}{2} \tag{52}$$

Example 15 *Suppose $N_m = 5$ and $N_f = 95$. So, the population size is $N = N_m + N_f = 100$. But, from (52), the effective population size is*

$$\begin{aligned} N_e &= \frac{1}{\frac{1}{4\times 5} + \frac{1}{4\times 95}} + \frac{1}{2} \\ &= 18.9899 + \frac{1}{2} \\ &= 19.4899 \end{aligned}$$

3.11.3 Distribution of Family Size

Let k_i be the number of gametes sampled from parent i. In the ideal population, each of the N parents is equally likely to contribute gametes to the next generation, and $2N$ gametes are sampled from the parents. So, in the ideal population, k_i is distributed as

$$k_i \sim \text{Binomial}(2N, \frac{1}{N}).$$

In most breeding populations, however, each parent is not equally likely to contribute gametes to the next generation. So, k_i will not have a Binomial distribution. We will now examine how the distribution of k_i affects the effective population size.

Let P_t denote the probability of self-fertilization. Then, the inbreeding coefficient in generation t can be written as

$$\begin{aligned} F_t &= P_t \frac{(1 + F_{t-1})}{2} + (1 - P_t) F_{t-1} \\ &= F_{t-1} + \frac{P_t}{2}(1 - F_{t-1}), \end{aligned} \tag{53}$$

and the rate of inbreeding becomes

$$\begin{aligned} \Delta_F &= \frac{F_t - F_{t-1}}{1 - F_{t-1}} \\ &= \frac{P_t}{2}. \end{aligned} \tag{54}$$

Thus from (43), the effective population size is the reciprocal of the probability that both alleles at a locus in generation t come from the same parent in generation $t - 1$.

$$\begin{aligned} N_e &= \frac{1}{2\Delta_{F_P}} \\ &= \frac{1}{P_t}. \end{aligned} \tag{55}$$

Note that in the ideal population, $P_t = \frac{1}{N}$. Thus, as expected, for the ideal population,

$$N_e = N.$$

We will now relate P_t to the distribution of k_i. Recall that P_t is the probability that both alleles at a locus in generation t come from the same parent in generation $t - 1$. Given that parent i transmits k_i gametes to the next generation, the number of ways in which you could choose two alleles from parent i is

$$\frac{k_i(k_i - 1)}{2},$$

and so the number of ways in which you could choose two alleles from the same parent is

$$\sum_i \frac{k_i(k_i - 1)}{2}.$$

The number of ways in which you could choose two alleles from $2N$ gametes is

$$\frac{2N(2N-1)}{2}.$$

So, conditional on a particular realization of the k_i's, the probability of self-fertilization is

$$\begin{aligned} \Pr(\text{selfing}|\boldsymbol{k}) &= \frac{\sum_i k_i(k_i-1)}{2N(2N-1)} \\ &= \frac{\sum_i k_i^2 - \sum_i k_i}{2N(2N-1)}, \end{aligned} \tag{56}$$

and the unconditional probability P_t of self-fertilization is

$$P_t = \frac{\sum_i \mathrm{E}(k_i^2) - \sum_i \mathrm{E}(k_i)}{2N(2N-1)}, \tag{57}$$

Note that in a population of constant size, $\mathrm{E}(k_i) = 2$. Now, using the notation

$$\bar{k} = \mathrm{E}(k_i)$$

and

$$V_k = \mathrm{Var}(k_i),$$

the first summation in the numerator of (57) is

$$\sum_i \mathrm{E}(k_i^2) = N(V_k + \bar{k}^2),$$

and the second summation in the numerator is

$$\sum_i \mathrm{E}(k_i) = 2N.$$

Now, P_t can be written as

$$\begin{aligned} P_t &= \frac{N(V_k+4) - 2N}{2N(2N-1)} \\ &= \frac{V_k+2}{4N-2}. \end{aligned} \tag{58}$$

Finally, using (55), the effective population size can be written in terms of V_k as

$$\begin{aligned} Ne &= \frac{1}{P_t} \\ &= \frac{4N-2}{V_k+2} \end{aligned} \tag{59}$$

Recall that in the ideal population

$$k_i \sim \text{Binomial}(2N, \frac{1}{N}).$$

Thus, the variance of k_i is

$$\begin{aligned} V_k &= 2N\frac{1}{N}(1-\frac{1}{N}) \\ &= 2(1-\frac{1}{N}), \end{aligned}$$

and substituting this V_k in (59), the effective population size for the ideal population is

$$\begin{aligned} N_e &= \frac{4N-2}{V_k+2} \\ &= \frac{4N-2}{2(1-\frac{1}{N})+2} \\ &= \frac{2N-1}{2-\frac{1}{N}} \\ &= \frac{N(2N-1)}{2N-1} \\ &= N. \end{aligned}$$

Suppose all parents contribute two gametes to the next generation, i.e., $k_i = 2$ for all i. Then, the variance of k_i is

$$V_k = 0,$$

and the efective population size becomes

$$\begin{aligned} N_e &= \frac{4N-2}{V_k+2} \\ &= \frac{4N-2}{2} \\ &= 2N-1 \\ &\approx 2N \end{aligned}$$

So, by making k_i a constant, the effective population size can be made almost twice the actual population size.
Now consider a population of bisexual organisms where the distribution of family size is not the same for males and females. Suppose the population consists of

$$N_m = \frac{N}{2}$$

males and

$$N_f = \frac{N}{2}$$

females. The effective population size for this population can be computed as

$$N_e \approx \frac{1}{Q_{t-1}} + \frac{1}{2},$$

which was derived in section 3.11.1. Here, Q_{t-1} is the probability that two alleles, say x and y, sampled from different individuals in generation $t-1$ originated in the same parent. Regardless of the distribution of family size, half the alleles in any generation originate in males. Thus, the probability that both x and y originate in males is $\frac{1}{4}$. The probability that both x and y originate in the same male parent can be computed as described below. Suppose the familiy size for male i is k_i, and

$$\sum_{i=1}^{\frac{N}{2}} k_i = N$$

Then, for a particular realization of the k_i's

$$\Pr(x, y \text{ are from same male} | \boldsymbol{k}) = \frac{\sum_{i=1}^{\frac{N}{2}} k_i(k_i-1)}{4N(N-1)},$$

and the unconditional probability is

$$\begin{aligned}\Pr(x, y \text{ are from same male}) &\approx \frac{\frac{N}{2}(V_m + 4) - N}{4N^2} \\ &\approx \frac{V_m + 2}{8N},\end{aligned}$$

where V_m is the variance of k_i and $\mathrm{E}(k_i) = 2$. Similarly, the probability that x and y are from the same male is

$$\Pr(x, y \text{ are from same female}) \approx \frac{V_f + 2}{8N},$$

where V_f is the variance of family size for females. Now, the probability that x and y are from the same parent is

$$Q_{t-1} \approx \frac{V_m + 2}{8N} + \frac{V_f + 2}{8N},$$

and the effective population size is

$$\begin{aligned}N_e &\approx \frac{1}{\frac{V_m+2}{8N} + \frac{V_f+2}{8N}} \\ &\approx \frac{8N}{V_m + V_f + 4}\end{aligned} \tag{60}$$

3.12 Equilibrium Between Drift and Mutation

Using the concept of effective population size in (37), in the absence of selection, mutation, or migration, the inbreeding coefficient in generation t can be written as

$$F_t = \frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1}. \tag{61}$$

This is the probability that the two alleles at a locus are IBD. However, if one of these alleles mutates, they will no longer be IBD. Therefore, the inbreeding coefficient when mutation is present is

$$F_t = [\frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1}](1 - u)^2 \tag{62}$$

At equilibrium,

$$F_E = F_t = F_{t-1},$$

and

$$F_E = [\frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_E](1-u)^2.$$

Solving for F_E in the above equation gives

$$\begin{aligned} F_E &= \frac{(1-u)^2}{2N_e - (2N_e - 1)(1-u)^2} \\ &= \frac{1 - 2u + u^2}{4N_e u + 1 - 2u - 2N_e u^2 + u^2} \\ &\approx \frac{1}{4N_e u + 1} \end{aligned} \tag{63}$$

3.13 Equilibrium Between Drift and Migration

Using the same approach as for mutation, ignoring the possibility of getting two migrant alleles that are IBD, the coefficient of inbreeding with migration is

$$F_t = [\frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1}](1-m)^2. \tag{64}$$

If m is very small, the equilibrium value of the inbreeding coefficient is

$$F_E \approx \frac{1}{4N_e m + 1} \tag{65}$$

3.14 Selection with Drift

3.14.1 Distribution of Gene Frequency

In section 3.10.2 we derived the distribution of gene frequency in a finite population, assuming no mutation, migration, or selection. Now we will see how this should be modified to account for selection. Recall that Y_t was defined as the number of "2"alleles in generation t. Then, frequency q_t of the "2" allele was defined as

$$q_t = \frac{Y_t}{2N}.$$

Because of this relationship, q_t and Y_t have the same shape. Thus, we will derive the distribution for Y_t.

At generation t, $2N$ alleles are sampled. Conditional on the frequency q_{t-1} in the previous generation, in the absence of selection, mutation, or migration, each sampled allele has a probability q_{t-1} of being a "2" allele. Thus, in generation t, the conditional distribution of the number of "2" alleles is:

$$(Y_t|Y_{t-1} = y_{t-1}) \sim \text{Binomial}(2N, q_{t-1} = \frac{y_{t-1}}{2N}).$$

However, if selection is present, an allele sampled in generation t will not have probability equal to the gene frequency in the previous generation. Selection will change this probability. Suppose q_{t-1} is the frequency in generation $t-1$. Then, as in (9), with selection, the probability of sampling a "2" allele in generation t is

$$q' = \frac{q_{t-1} - sq_{t-1}^2 - hs(1-q_{t-1})q_{t-1}}{1 - 2hs(1-q_{t-1})q_{t-1} - sq_{t-1}^2},$$

and thus with selection, in generation t, the conditional distribution of the number of "2" alleles is:

$$(Y_t|Y_{t-1} = y_{t-1}) \sim \text{Binomial}(2N, q').$$

The unconditional distribution of the number of "2" alleles is given by (click here for plot)

$$\Pr(Y_t = y_t) = \sum_{y_{t-1}=0}^{2N} \Pr(Y_{t-1} = y_{t-1}) \Pr(Y_t = y_t|Y_{t-1} = y_{t-1}). \quad (66)$$

3.14.2 Approximation to Probability of Fixation

Consider selection for an additive trait in an ideal population of size N. Denote the difference between gene frequencies between generations t and $t+1$ by

$$\Delta_t = p_{t+1} - p_t.$$

Formula (12) in section 3.7 gives the change in gene frequency due to selection in an infinite population. In a finite population, this formula gives the expected change in gene frequency due to selection. Approximating $\bar{W}$ by 1.0 and taking $h = \frac{1}{2}$, in 12, the conditional expectation of Δ_t given p_t becomes

$$\text{E}(\Delta_t|p_t) \approx \frac{1}{2}sp_t(1-p_t),$$

and the unconditional expectation is

$$\mathrm{E}(\Delta_t) \approx \frac{1}{2} s \mathrm{E}[p_t(1-p_t)]. \tag{67}$$

The expected value on the right can be written as

$$\mathrm{E}[p_t(1-p_t)] = \mathrm{E}(p_t) - [\mathrm{E}(p_t)]^2 - \mathrm{Var}(p_t). \tag{68}$$

When Ns is small, ignoring selection, this can be approximated in terms of the gene frequency p in generation 0 as

$$\begin{aligned} \mathrm{E}[p_t(1-p_t)] &= p - p^2 - \mathrm{Var}(p_t) \\ &= p(1-p) - \mathrm{Var}(p_t) \\ &= p(1-p)(1-\frac{1}{2N})^t, \end{aligned} \tag{69}$$

because from (31),

$$\mathrm{Var}(p_t) = p(1-p)[1-(1-\frac{1}{2N})^t]$$

Substituting (69) in (67) gives

$$\mathrm{E}(\Delta_t) \approx \frac{1}{2} sp(1-p)(1-\frac{1}{2N})^t \tag{70}$$

As t goes to infinity, the frequency of the favorable allele is 0 or 1, and thus, the expected limiting gene frequency is equal to the probability of fixation. Given frequency p for the favorable allele in generation 0, the expected value of the limiting gene frequency is:

$$\begin{aligned} u(p) &= p + \sum_{t=0}^{\infty} \mathrm{E}(\Delta_t) \\ &\approx p + \frac{1}{2} sp(1-p) \sum_{t=0}^{\infty} (1-\frac{1}{2N})^t \\ &\approx p + Nsp(1-p), \end{aligned} \tag{71}$$

which is also the probability of fixation. For an arbitrary population, the probability of fixation is

$$u(p) \approx p + N_e sp(1-p), \tag{72}$$

where N_e is the effective population size. Thus when $N_e s$ is small, the limiting response to selection is

$$\begin{aligned} \Delta_{p_\infty} &= u(p) - p \\ &\approx N_e sp(1-p), \end{aligned} \tag{73}$$

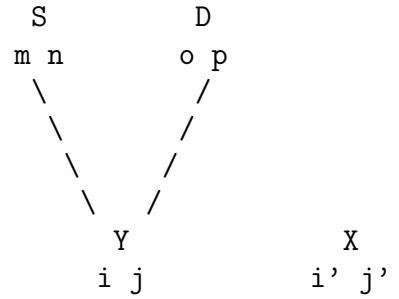
which is $2N_e$ times $\frac{1}{2}sp(1-p)$, the initial response to selection.

3.15 Inbreeding with Pedigree

When the pedigree for an individual is available, the inbreeding specific to that individual can be computed. The coefficient of coancestry will be used to this. Suppose X and Y are the parents of Z. Then, the inbreeding coefficient F_Z for individual Z is

$$F_Z = r_{XY},$$

where r_{XY} is the coefficient of coancestry between X and Y. We will now develop a recursive formula to compute r_{XY}, using pedigree information.
Suppose X is not a direct descendant of Y, and let S and D be the father and mother of Y. The alleles of S, D, X, and Y are labelled as shown in the following diagram.



Recall that r_{XY} is the probability that a random allele from X is IBD to a random allele from Y. The random allele from Y can be i or j with equal probability, and the random allele from X can be i' or j' with equal probability. So,

$$r_{XY} = \frac{1}{4}[\Pr(i \equiv i') + \Pr(i \equiv j') + \Pr(j \equiv i') + \Pr(j \equiv j')] \quad (74)$$

Let i be the paternal allele of Y. Then, i is either m or n with equal probability. So,

$$\Pr(i \equiv i') = \frac{1}{2}[\Pr(m \equiv i') + \Pr(n \equiv i')]$$

Similarly,

$$\Pr(i \equiv j') = \frac{1}{2}[\Pr(m \equiv j') + \Pr(n \equiv j')],$$

$$\Pr(j \equiv i') = \frac{1}{2}[\Pr(o \equiv i') + \Pr(p \equiv i')],$$

and

$$\Pr(j \equiv j') = \frac{1}{2}[\Pr(o \equiv j') + \Pr(p \equiv j')]$$

Substituting the above in (74) gives

$$\begin{aligned} r_{XY} = \frac{1}{4}\{ & \\ & \frac{1}{2}[\Pr(m \equiv i') + \Pr(n \equiv i') + \Pr(m \equiv j') + \Pr(n \equiv j')] \\ & + \frac{1}{2}[\Pr(o \equiv i') + \Pr(p \equiv i') + \Pr(o \equiv j') + \Pr(p \equiv j')] \\ \} & \end{aligned} \quad (75)$$

This can be written as

$$\begin{aligned} r_{XY} = \frac{1}{2}\{ & \\ & \frac{1}{4}[\Pr(m \equiv i') + \Pr(n \equiv i') + \Pr(m \equiv j') + \Pr(n \equiv j')] \\ & + \frac{1}{4}[\Pr(o \equiv i') + \Pr(p \equiv i') + \Pr(o \equiv j') + \Pr(p \equiv j')] \\ & \} \\ & = \frac{1}{2}(r_{XS} + r_{XD}) \end{aligned} \tag{76}$$

Thus, the coefficient of coancestry between X and Y is the average coancestry between X and the parents of Y. Note that in order to compute coancestry by (76), the condition that X cannot be a descendant of Y must be true. Choosing Y to be the younger of the two individuals will ensure that this condition is always true.

3.16 Tabular Method to Compute Coancestry

The following rules can be used to compute the coancestry between each pair of individuals in a pedigree.

1. Number individuals such that parents precede offspring.
2. For founders, enter $\frac{1}{2}$ on the diagonals and 0 on the off-diagonals.
3. For non-founder individual i,
   (a) calculate row element 1 to $i-1$ as the average of the parental row elements,
   (b) set diagonal element to
   $$\frac{1}{2}(1 + r_{SD}),$$
   where S and D are the parents of i.
4. Complete column i by symmetry

3.17 Regular Systems of Inbreeding

3.17.1 Self-Fertilization

$$\begin{aligned} F_X &= r_{AA} \\ &= \frac{1}{2}(1 + F_A) \end{aligned}$$

At generation t, the inbreeding coeficient is

$$F_t = \frac{1}{2}(1 + F_{t-1})$$

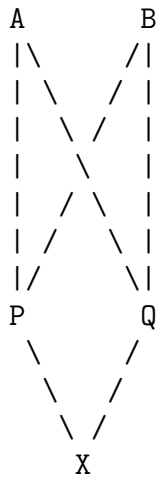
3.17.2 Parent-Offspring Mating

$$\begin{aligned} F_x &= r_{AP} \\ &= \frac{1}{2}(r_{AA} + r_{AB}) \\ &= \frac{1}{2}[\frac{1}{2}(1 + F_A) + F_P] \\ &= \frac{1}{4}(1 + F_A + 2F_P) \end{aligned}$$

At generation t,

$$F_t = \frac{1}{4}(1 + 2F_{t-1} + F_{t-2})$$

3.17.3 Fullsib Mating



$$
\begin{aligned}
F_X &= r_{PQ} \\
&= \frac{1}{2}(r_{PA} + r_{PB}) \\
&= \frac{1}{2}[\frac{1}{2}(r_{AA} + r_{AB}) + \frac{1}{2}(r_{AB} + r_{BB})] \\
&= \frac{1}{4}[\frac{1}{2}(1 + F_A) + \frac{1}{2}(1 + F_B) + 2F_P]
\end{aligned}
$$

At generation t,

$$F_t = \frac{1}{4}(1 + 2F_{t-1} + F_{t-2})$$

t	F	Δ_F
1	0.250	0.250
2	0.375	0.167
3	0.500	0.200
4	0.594	0.188
$\vdots$	$\vdots$	$\vdots$
20	0.986	0.191

4 Multi-Locus Inheritance

4.1 Genotypic Value

Genotypic value for genotype T is defined as

$$G = \mathrm{E}(P|T)$$

where P is the phenotype. So, can write P as

$$P = G + E$$

where $E = P - G$.
From property 1 of regression (page 14), $\mathrm{E}(G) = \mathrm{E}(P)$,
From property 2 of regression (page 15) , $\mathrm{E}(E) = 0$
From property 3 of regression (page 15), $\mathrm{Cov}(G, E) = 0$.

4.2 Resemblance between Relatives

Resemblance between x and y measured by:

$$\text{Cov}(P_x, P_y)$$

To measure genetic resemblance, phenotypic value is modeled as:

$$P = G + E$$

Then,

$$\begin{aligned}\text{Cov}(P_x, P_y) &= \text{Cov}(G_x, G_y) + \text{Cov}(E_x, E_y) \\ &= \text{Cov}(G_x, G_y)\end{aligned}$$

if $\text{Cov}(E_x, E_y) = 0$

4.3 Multifactorial Model

The covariance between relatives is due to IBD the alleles they share. Thus, to derive the covariance between relatives, it is convenient to model the genotypic value as the sum of the effects due to the alleles and their interactions.

4.3.1 Notation and Assumptions

one locus with two alleles a_1 and a_2
paternal allele = A_i
maternal allele = A_j
$\Pr(A_i = a_1) = p$
$\Pr(A_i = a_2) = 1 - p = q$
Hardy-Weinberg equilibrium

Genotype (T)	Genotypic Value (G)	$\Pr(T)$
a_1a_1	a	p^2
a_1a_2	d	$2pq$
a_2a_2	$-a$	q^2

a is the genotypic value for genotypes a_1a_1
d is the genotypic value for genotypes a_1a_2
Both, a and d are relative to the midpoint between the genotypic values for two homozygous genotypes.

4.3.2 Genotypic Mean

$$\begin{aligned}\mu &= \mathrm{E}(G) \\ &= a(p^2 - q^2) + 2dpq \\ &= a(p-q)(p+q) + 2dpq \\ &= a(p-q) + 2dpq\end{aligned}$$

4.3.3 Model for Genotypic value: Step 1

$$G = \mu + (G - \mu)$$

μ does not contribute to the covariance between relatives
Easier to work with $(G - \mu)$ because it has null mean

4.3.4 Model for Genotypic value: Step 2

$$\begin{aligned}(G - \mu) &= \mathrm{E}[(G - \mu)|A_i] + \epsilon \\ &= \alpha_i + \epsilon\end{aligned}$$

where

$$\epsilon = (G - \mu) - \alpha_i$$

From property 2 of regression (page 15),

$$\mathrm{E}(\epsilon) = 0$$

α_i is the regression of $(G - \mu)$ on A_i
It is the component of the genotypic value associated with A_i and is called the average effect of allele a_i
From property 1 of regression (page 14),

$$\begin{aligned}\mathrm{E}(\alpha_i) &= \mathrm{E}(G - \mu) \\ &= 0\end{aligned}$$

4.3.5 Average Effect of Allele a_1

$$\begin{aligned}
\alpha_1 &= \mathrm{E}[(G-\mu)|A_i = a_1] \\
&= \mathrm{E}(G|A_i = a_1) - \mu \\
&= pa + qd - [a(p-q) + 2dpq] \\
&= pa + qd - pa + qa - 2pqd \\
&= qa + qd(1-2p) \\
&= q[a + d(1-2p)] \\
&= q[a + d(p+q-2p)] \\
&= q[a + d(q-p)]
\end{aligned}$$

4.3.6 Average Effect of Allele a_2

$$\begin{aligned}
\alpha_2 &= \mathrm{E}[(G-\mu)|A_i = a_2] \\
&= \mathrm{E}(G|A_i = a_2) - \mu \\
&= pd - qa - [a(p-q) + 2dpq] \\
&= pd - qa - pa + qa - 2pqd \\
&= -pa + pd(1-2q) \\
&= -p[a + d(2q-1)] \\
&= -p[a + d(2q-p-q)] \\
&= -p[a + d(q-p)]
\end{aligned}$$

4.3.7 Model for Genotypic value: Step 3

$$\begin{aligned}
\epsilon &= \mathrm{E}(\epsilon|A_j) + \delta_{ij} \\
&= \mathrm{E}[(G-\mu-\alpha_i)|A_j] + \delta_{ij} \\
&= \mathrm{E}[(G-\mu)|A_j] - \mathrm{E}(\alpha_i|A_j) + \delta_{ij}
\end{aligned}$$

Under Hardy-Weinberg equilibrium, A_i and A_j are sampled independently. So,

$$\begin{aligned}
\epsilon &= \mathrm{E}[(G-\mu)|A_j] - E(\alpha_i) + \delta_{ij} \\
&= \mathrm{E}[(G-\mu)|A_j] + \delta_{ij} \\
&= \alpha_j + \delta_{ij}
\end{aligned}$$

From property 2 of regression (page 15),

$$\mathrm{E}(\delta_{ij}) = 0$$

α_j is the regression of $(G - \mu)$ on A_j
It is the component of the genotypic value associated with A_j.
Because A_i and A_j are sampled from the same population,

$$\mathrm{E}[(G - \mu)|A_j = a_1] = \alpha_1$$

and

$$\mathrm{E}[(G - \mu)|A_j = a_2] = \alpha_2$$

Further,

$$\mathrm{E}(\alpha_j) = 0$$

It can also be shown that $\mathrm{E}(\delta_{ij}|A_i) = 0$.

4.3.8 Model for Genotypic value

$$G = \mu + \alpha_i + \alpha_j + \delta_{ij}$$

Recall that from regression theory,

$$\begin{aligned}\mathrm{Cov}(\alpha_i, \epsilon) &= \mathrm{Cov}(\alpha_i, \alpha_j + \delta_{ij}) \\ &= \mathrm{Cov}(\alpha_i, \alpha_j) + \mathrm{Cov}(\alpha_i, \delta_{ij}) \\ &= 0\end{aligned}$$

and

$$\mathrm{Cov}(\alpha_j, \delta_{ij}) = 0$$

Further, from Hardy-Weinberg equilibrium,

$$\mathrm{Cov}(\alpha_i, \alpha_j) = 0$$

So,

$$\mathrm{Cov}(\alpha_i, \delta_{ij}) = 0$$

4.4 Genotypic Variance

$$\mathrm{Var}(G) = \mathrm{Var}(\alpha_i) + \mathrm{Var}(\alpha_j) + \mathrm{Var}(\delta_{ij})$$

where
$\mathrm{Var}(\alpha_i) + \mathrm{Var}(\alpha_j)$ is the additive variance: V_A
$\mathrm{Var}(\delta_{ij})$ is the dominance variance: V_D

4.5 Covariance between Relatives

Let

$$G_x = \mu + \alpha_i + \alpha_j + \delta_{ij}$$

and

$$G_y = \mu + \alpha_{i'} + \alpha_{j'} + \delta_{i'j'}$$

Then,

$$\begin{aligned}\mathrm{Cov}(G_x, G_y) &= \mathrm{Cov}(\alpha_i, \alpha_{i'}) + \mathrm{Cov}(\alpha_i, \alpha_{j'}) \\ &\quad + \mathrm{Cov}(\alpha_j, \alpha_{i'}) + \mathrm{Cov}(\alpha_j, \alpha_{j'}) \\ &\quad + \mathrm{Cov}(\delta_{ij}, \delta_{i'j'})\end{aligned}$$

4.5.1 IBD Alleles

Two alleles are said to be identical by descent (IBD) if one is a copy of the other or both are copies of a common allele.
$A_i \equiv A_j$ denotes that A_i and A_j are IBD
$A_i \not\equiv A_j$ denotes that A_i and A_j are not IBD.
Resemblance between relatives results from relatives sharing alleles that are IBD

4.5.2 Additive Covariance

Let Z denote the IBD state of alleles A_i and $A_{i'}$.
Thus, Z has two states: $A_i \equiv Ai'$ and $A_i \not\equiv A_{i'}$
Then,

$$\begin{aligned}\mathrm{Cov}(\alpha_i, \alpha_{i'}) &= \mathrm{E}(\alpha_i \alpha_{i'}) - \mathrm{E}(\alpha_i)\mathrm{E}(\alpha_{i'}) \\ &= \mathop{\mathrm{E}}_{Z}[\mathrm{E}(\alpha_i \alpha_{i'}|Z)] \\ &= \mathrm{E}(\alpha_i \alpha_{i'}|A_i \equiv A_{i'}) \Pr(A_i \equiv Ai') \\ &+ \mathrm{E}(\alpha_i \alpha_{i'}|A_i \not\equiv A_{i'}) \Pr(A_i \not\equiv A_{i'}) \\ &= \mathrm{E}(\alpha_i^2) \Pr(A_i \equiv A_{i'}) \\ &+ \mathrm{E}(\alpha_i)\mathrm{E}(\alpha_{i'}) \Pr(A_i \not\equiv A_{i'}) \\ &= \mathrm{Var}(\alpha_i) \Pr(A_i \equiv A_{i'})\end{aligned}$$

Similarly,

$$\mathrm{Cov}(\alpha_i, \alpha_{j'}) = \mathrm{Var}(\alpha_i)\Pr(A_i \equiv A_{j'})$$

$$\mathrm{Cov}(\alpha_j, \alpha_{i'}) = \mathrm{Var}(\alpha_j)\Pr(A_j \equiv A_{i'})$$

$$\mathrm{Cov}(\alpha_j, \alpha_{j'}) = \mathrm{Var}(\alpha_j)\Pr(A_j \equiv A_{j'})$$

Now,

$$\begin{aligned}\mathrm{Cov}(G_x, G_y) = &\mathrm{Var}(\alpha_i)(\phi_{ii'} + \phi_{ij'}) \\ &+ \mathrm{Var}(\alpha_j)(\phi_{ji'} + \phi_{jj'}) \\ &+ \mathrm{Cov}(\delta_{ij}, \delta_{i'j'})\end{aligned}$$

where $\phi_{ii'}$, for example, denotes $\Pr(A_i \equiv A_{i'})$.
But, $\mathrm{Var}(\alpha_i) = \mathrm{Var}(\alpha_j) = V_A/2$. So,

$$\mathrm{Cov}(G_x, G_y) = a_{xy}V_A + \mathrm{Cov}(\delta_{ij}, \delta_{i'j'})$$

where

$$a_{xy} = \frac{(\phi_{ii'} + \phi_{ij'} + \phi_{ji'} + \phi_{jj'})}{2}$$

a_{xy} is called the additive relationship coefficient

4.5.3 Computing a_{xy}

Suppose:
y is not a descendant of x
s is the father of i with alleles A_m and A_n
d is the mother of i with alleles A_o and A_p
Then,

$$\phi_{ii'} = \frac{1}{2}(\phi_{mi'} + \phi_{ni'})$$

$$\phi_{ij'} = \frac{1}{2}(\phi_{mj'} + \phi_{nj'})$$

$$\phi_{ji'} = \frac{1}{2}(\phi_{oi'} + \phi_{pi'})$$

$$\phi_{jj'} = \frac{1}{2}(\phi_{oj'} + \phi_{pj'})$$

and,

$$\begin{aligned} a_{xy} &= \frac{1}{2}(\phi_{ii'} + \phi_{ij'} + \phi_{ji'} + \phi_{jj'}) \\ &= \frac{1}{2}[\frac{1}{2}(\phi_{mi'} + \phi_{ni'} + \phi_{mj'} + \phi_{nj'}) \\ &\quad + \frac{1}{2}(\phi_{oi'} + \phi_{pi'} + \phi_{oj'} + \phi_{pj'})] \\ &= \frac{1}{2}(a_{sy} + a_{dy}) \end{aligned}$$

4.5.4 Additive Relationship Matrix

Example 16

Pedigree:

Individual	Father	Mother
1	0	0
2	0	0
3	1	2
4	1	2
5	0	0
6	4	5
7	4	5

Relationship Matrix:

1	0	0.5	0.5	0	0.25	0.25
0	1	0.5	0.5	0	0.25	0.25
0.5	0.5	1	0.5	0	0.25	0.25
0.5	0.5	0.5	1	0	0.5	0.5
0	0	0	0	1	0.5	0.5
0.25	0.25	0.25	0.5	0.5	1	0.5
0.25	0.25	0.25	0.5	0.5	0.5	1

4.5.5 Tabular Method

Number individuals by birth order

For a base individual, i, set to zero row elements 1 to i-1
For a non-base individual, i, calculate row elements 1 to i-1 as the average of the parental row elements
Set diagonal to 1
Complete column by symmetry

4.5.6 Dominance Covariance

Let

$$W = 1 \qquad \text{if} \begin{cases} A_i \equiv A_{i'}, A_j \equiv A_{j'} \\ \text{or} \\ A_i \equiv A_{j'}, A_j \equiv A_{i'} \end{cases}$$

and otherwise

$$W = 2$$

Then,

$$\begin{aligned} \mathrm{Cov}(\delta_{ij}, \delta_{i'j'}) &= \mathrm{E}(\delta_{ij}\delta_{i'j'}) - \mathrm{E}(\delta_{ij})\mathrm{E}\delta_{i'j'}) \\ &= \underset{W}{\mathrm{E}}[\mathrm{E}(\delta_{ij}\delta_{i'j'})|W] \\ &= \mathrm{E}(\delta_{ij}\delta_{i'j'})|W = 1)\,\Pr(W = 1) \\ &+ \mathrm{E}(\delta_{ij}\delta_{i'j'})|W = 2)\,\Pr(W = 2) \\ &= \mathrm{E}(\delta_{ij}^2)\,\Pr(W = 1) \\ &= \mathrm{Var}(\delta_{ij})\,\Pr(W = 1) \end{aligned}$$

4.5.7 Genotypic Covariance

Finally,

$$\mathrm{Cov}(G_x, G_y) = a_{xy}V_A + u_{xy}V_D$$

where

$$u_{xy} = \Pr(W = 1)$$

In general,
V_A: sum of the additive variances over all loci
V_D: sum of the dominance variances over all loci

4.5.8 Computing u_{xy}

Let s and d be the parents of x and s' and d' be the parents of y. In the absence of inbreeding,

$$u_{xy} = (\phi_{ii'}\phi_{jj'} + \phi_{ij'}\phi_{ji'})$$

Note that i is a random allele from s and i' is a random allele from s'. So,

$$\phi_{ii'} = r_{ss'}$$

where $r_{ss'}$ is the coefficient of coancestry between s and s'. Similarly,

$$\phi_{jj'} = r_{dd'}$$

$$\phi_{ij'} = r_{sd'}$$

and

$$\phi_{ji'} = r_{ds'}$$

Finally,

$$u_{xy} = \frac{1}{4}(a_{ss'}a_{dd'} + a_{sd'}a_{ds'})$$

because, $r_{ij} = \frac{1}{2}a_{ij}$

4.5.9 Relationship Coefficients

Relationship	a_{xy}	u_{xy}
Identical twins	1	1
Parent-offspring	0.5	0
Grandparent-offspring	0.25	0
Full sibs	0.5	0.25
Half sibs	0.25	0
Uncle(Aunt)-nephew(niece)	0.25	0
First cousins	$\frac{1}{8}$	0
Double first cousins	$\frac{1}{4}$	$\frac{1}{16}$

4.6 Covariance Between Traits

Assume covariance between traits is due to pleiotropy
Notation:
G_x^1 is the genotypic value for trait 1 in x
G_x^2 is the genotypic value for trait 2 in x
A_x^1 is the sum of the additive effects for trait 1 in x
A_x^2 is the sum of the additive effects for trait 2 in x
D_x^1 is the dominance effect for trait 1 in x
D_x^2 is the dominance effect for trait 2 in x
$\mathrm{Cov}(A_x^1, A_x^2) = C_A^{12}$
$\mathrm{Cov}(D_x^1, D_x^2) = C_D^{12}$
Then,

$$\mathrm{Cov}(G_x^1, G_x^2) = C_A^{12} + C_D^{12}$$

and

$$\mathrm{Cov}(G_x^1, G_y^2) = a_{xy} C_A^{12} + u_{xy} C_D^{12}$$

4.7 Response to Selection

We will use linear regression to study response to selection.

4.7.1 Linear Regression

The linear regression of X on Y is:

$$\hat{X} = \mathrm{E}(X) + \beta[Y - \mathrm{E}(Y)]$$

where

$$\beta = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}$$

It is easy to see that $\mathrm{E}(\hat{X}) = \mathrm{E}(X)$:

$$\begin{aligned} \mathrm{E}(\hat{X}) &= \mathrm{E}(X) + \beta[\mathrm{E}(Y) - \mathrm{E}(Y)] \\ &= \mathrm{E}(X) \end{aligned}$$

Further, if X and Y have a bivariate normal distribution, can show that

$$\mathrm{E}(X|Y) = \hat{X}$$

To show this, write

$$X = \hat{X} + \epsilon$$

where

$$\begin{aligned}\epsilon &= X - \hat{X} \\ &= X - \{\mathrm{E}(X) + \beta[Y - \mathrm{E}(Y)]\}\end{aligned}$$

Then,

$$\begin{aligned}\mathrm{E}(\epsilon) &= \mathrm{E}(X) - \mathrm{E}(\hat{X}) \\ &= \mathrm{E}(X) - \mathrm{E}(X) \\ &= 0\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(\epsilon, Y) &= \mathrm{Cov}(X - \{\mathrm{E}(X) + \beta[Y - \mathrm{E}(Y)]\}, Y) \\ &= \mathrm{Cov}(X, Y) - \beta\mathrm{Cov}(Y, Y) \\ &= \mathrm{Cov}(X, Y) - \beta\mathrm{Var}(Y) \\ &= 0\end{aligned}$$

because $\beta\mathrm{Var}(Y) = \mathrm{Cov}(X, Y)$. Further, because Y and ϵ are normally distributed, the null covariance implies they are also independent. Thus, under normality,

$$\begin{aligned}\mathrm{E}(X|Y) &= \mathrm{E}(\hat{X}|Y) + \mathrm{E}(\epsilon|Y) \\ &= \hat{X} + \mathrm{E}(\epsilon) \\ &= \hat{X}\end{aligned}$$

Further, under normality,

$$\begin{aligned}\mathrm{Var}(X|Y) &= \mathrm{Var}(\hat{X}|Y) + \mathrm{Var}(\epsilon|Y) \\ &= \mathrm{Var}(\epsilon) \\ &= \mathrm{Var}(X - \{\mathrm{E}(X) + \beta[Y - \mathrm{E}(Y)]\}) \\ &= \mathrm{Var}(X - \beta Y) \\ &= \mathrm{Var}(X) - 2\beta\mathrm{Cov}(X, Y) + \beta^2\mathrm{Var}(Y) \\ &= \mathrm{Var}(X) - \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(Y)}\end{aligned}$$

because $\beta\mathrm{Var}(Y) = \mathrm{Cov}(X, Y)$.

4.7.2 Truncation Selection

Suppose $Y \sim N(\mu_Y, V_Y)$ The mean and variance of Y given truncation selection are:

$$\mathrm{E}(Y|Y > t) = \mu_Y + V_Y^{1/2} i \tag{77}$$

where

$$i = \frac{f(s)}{p}$$

$f(s)$ is the standard normal density function

$$s = \frac{t - \mu_Y}{V_Y^{1/2}}$$

$$p = \Pr(Y > t)$$

and

$$\mathrm{Var}(Y|Y > t) = V_Y[1 - i(i - s)] \tag{78}$$

To prove the above, we first derive the mean and variance for a standard normal variable given truncation selection.
Let $Z \sim N(0, 1)$. The density function of Z is:

$$f(z) = \sqrt{\frac{1}{2\pi}} e^{-\frac{1}{2}z^2}$$

The density function for Z given truncation selection is

$$f(z|z > s) = f(z)/p$$

From the definition of the mean,

$$\begin{aligned}
\mathrm{E}(Z|Z > s) &= \frac{1}{p}\int_s^\infty z f(z) dz \\
&= \frac{1}{p}[-f(z)]_s^\infty \\
&= \frac{f(s)}{p} \\
&= i
\end{aligned}$$

because the first derivative of $f(z)$ with respect to z is:

$$\begin{aligned}\frac{d}{dz}f(z) &= \sqrt{\frac{1}{2\pi}}e^{-\frac{1}{2}z^2}(-z) \\ &= -zf(z)\end{aligned}$$

Now, to compute the variance of Z given selection, consider the following identity:

$$\begin{aligned}\frac{d}{dz}zf(z) &= f(z) + z\frac{d}{dz}f(z) \\ &= f(z) - z^2 f(z)\end{aligned}$$

Integrating both sides from s to ∞ gives

$$zf(z)]_s^\infty = \int_s^\infty f(z)dz - \int_s^\infty z^2 f(z)dz$$

Upon rearranging this gives:

$$\begin{aligned}\int_s^\infty z^2 f(z)dz &= \int_s^\infty f(z)dz - zf(z)]_s^\infty \\ \frac{1}{p}\int_s^\infty z^2 f(z)dz &= \frac{1}{p}\int_s^\infty f(z)dz + \frac{f(s)}{p}s \\ &= 1 + is\end{aligned}$$

So,

$$\begin{aligned}\mathrm{Var}(Z|Z > s) &= 1 + is - i^2 \\ &= 1 - i(i - s)\end{aligned} \tag{79}$$

As shown below, the results for Y follow from the fact that

$$\mu_Y + V_Y^{1/2}Z \sim N(\mu_Y, V_Y)$$

Let

$$Y = \mu_Y + V_Y^{1/2}Z,$$

Then, the condition

$$Y > t$$

is equivalent to

$$\begin{aligned} \mu_Y + V_Y^{1/2} Z &> t \\ V_Y^{1/2} Z &> t - \mu_Y \\ Z &> \frac{t - \mu_Y}{V_Y^{1/2}} \\ Z &> s \end{aligned}$$

Now,

$$\begin{aligned} \mathrm{E}(Y|Y > t) &= \mathrm{E}(\mu_Y + V_Y^{1/2} Z | Z > s) \\ &= \mu_Y + V_Y^{1/2} i, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(Y|Y > t) &= \mathrm{Var}(\mu_Y + V_Y^{1/2} Z | Z > s) \\ &= V_Y[1 - i(i - s)] \end{aligned}$$

4.7.3 Correlated Response to Selection

Suppose:
X and Y are bivariate normal
$\mathrm{E}(X) = \mu_X$
$\mathrm{E}(Y) = \mu_Y$
$\mathrm{Var}(X) = V_X$
$\mathrm{Var}(Y) = V_Y$
$\mathrm{Cov}(X, Y) = C_{XY}$
Using the double expectation theorem,

$$\begin{aligned} \mathrm{E}(X|Y > t) &= \mathrm{E}[\mathrm{E}(X|Y)|Y > t] \\ &= \mathrm{E}[\mu_X + \beta(Y - \mu_Y)|Y > t] \\ &= \mu_X + \beta[\mathrm{E}(Y|Y > t) - \mu_Y] \\ &= \mu_X + \beta V_Y^{1/2} i \end{aligned}$$

because $\mathrm{E}(Y|Y > t) = \mu_Y + V_Y^{1/2} i$ (page 68).

Using the identity for the variance given on page 11,

$$\begin{aligned}
\mathrm{Var}(X|Y>t) &= \mathrm{E}[\mathrm{Var}(X|Y)|Y>t] + \mathrm{Var}[\mathrm{E}(X|Y)|Y>t] \\
&= \mathrm{E}[(V_X - \beta C_{XY})|Y>t] + \mathrm{Var}\{[\mu_X + \beta(Y-\mu_Y)|Y>t]\} \\
&= V_X - \beta C_{XY} + \beta^2 \mathrm{Var}(Y|Y>t) \\
&= V_X - \beta^2 V_Y + \beta^2 \mathrm{Var}(Y|Y>t) \\
&= V_X - \beta^2 [V_Y - \mathrm{Var}(Y|Y>t)] \\
&= V_X - \beta^2 V_Y i(i-s)
\end{aligned}$$

because $\mathrm{Var}(Y|Y>t) = V_Y[1 - i(i-s)]$ (page 68).

4.7.4 Regression of Offspring on Mid-parent

Let P_x, P_s and P_d denote the phenotypic values of an individual and its parents. Then,

$$\begin{aligned}
\mathrm{Cov}(P_x, \frac{P_s + P_d}{2}) &= \frac{1}{2}[\mathrm{Cov}(P_x, P_s) + \mathrm{Cov}(P_x, P_d)] \\
&= \frac{1}{2}[\frac{1}{2}V_A + \frac{1}{2}V_A] \\
&= \frac{1}{2}V_A
\end{aligned}$$

and the variance of the mid-parent value is:

$$\mathrm{Var}(\frac{P_s + P_d}{2}) = \frac{1}{2}V_P$$

Thus, under normality, the regression of offspring on mid-parent is

$$\mathrm{E}(P_x|\frac{P_s + P_d}{2}) = \mu + \frac{V_A}{V_P}(\frac{P_s + P_d}{2} - \mu)$$

The slope of this regression line is:

$$h^2 = \frac{V_A}{V_P}$$

and is called the heritability.

4.7.5 Response To Selection: Mean and Variance

Generation 0: $\mathrm{E}(P) = \mu_{P_0}$

$\mathrm{E}(A) = \mu_{A_0} = 0$

$\mathrm{E}(D) = \mu_{D_0} = 0$

$\mathrm{Var}(P) = V_{P_0}$

$\mathrm{Var}(A) = V_{A_0}$

$\mathrm{Var}(D) = V_{D_0}$

Generation 1: Note that the phenotypic value of a parent is uncorrelated with the dominance effect and environmental deviation of an offspring. Thus, under normality, the phenotypic value of the parent is independent of the dominance effect and the environmental deviation of the offspring. So, selection of parents has an effect only on the additive effect of the offspring. To study the effect of truncation selection on P_s and P_d, we model A_x as

$$A_x = \tfrac{1}{2}A_s + \tfrac{1}{2}A_d + \epsilon_x$$

Computing the covariance of P_s with both sides of the model for A_x gives

$$\begin{aligned}\mathrm{Cov}(P_s, A_x) &= \tfrac{1}{2}\mathrm{Cov}(P_s, A_s) + \tfrac{1}{2}\mathrm{Cov}(P_s, A_d) + \mathrm{Cov}(P_s, \epsilon_x) \\ \tfrac{1}{2}V_A &= \tfrac{1}{2}V_A + \mathrm{Cov}(P_s, \epsilon_x)\end{aligned}$$

because we assume parents are unrelated. This implies that $\mathrm{Cov}(P_s, \epsilon_x) = 0$. Under normality, $\mathrm{Cov}(P_s, \epsilon_x) = 0$ implies that P_s is independent of ϵ_x. Similarly, P_d is also independent of ϵ_x.
Now, the mean of A_x given selection of parents in generation 0 is

$$\begin{aligned}\mathrm{E}(A_x|\mathrm{Sel}_0) &= \tfrac{1}{2}\mathrm{E}(A_s|P_s > t) + \tfrac{1}{2}\mathrm{E}(A_d|P_d > t) \\ &= \frac{V_{A_0}}{V_{P_0}} V_{P_0}^{1/2} i \\ &= h_0^2 V_{P_0}^{1/2} i\end{aligned}$$

where

$$h_0^2 = \frac{V_{A_0}}{V_{P_0}}$$

The variance of A_x given selection of parents in generation 0 is

$$\begin{aligned}\mathrm{Var}(A_x|\mathrm{Sel}_0) &= \frac{1}{4}\mathrm{Var}(A_s|P_s > t) + \frac{1}{4}\mathrm{Var}(A_d|P_d > t) + \mathrm{Var}(\epsilon_x) \\ &= \tfrac{1}{2}[V_{A_0} - \frac{V_{A_0}^2}{V_{P_0}^2} V_{P_0} i(i-s)] + \tfrac{1}{2}V_{A_0} \\ &= \tfrac{1}{2}V_{A_0}[1 - h_0^2 i(i-s)] + \tfrac{1}{2}V_{A_0}\end{aligned}$$

Generation 2: The mean of A_x given selection of parents in generations 0 and 1 is

$$\mathrm{E}(A_x|\mathrm{Sel}_1) = h_0^2 V_{P_0}^{1/2} i + h_1^2 V_{P_1}^{1/2} i$$

and the variance of A_x given selection of parents in generations 0 and 1 is

$$\mathrm{Var}(A_x|\mathrm{Sel}_1) = \tfrac{1}{2}V_{A_1}[1 - h_1^2 i(i-s)] + \tfrac{1}{2}V_{A_0}$$

Generation t: In generation t, the mean of A_x given selection of parents for t generations is

$$\mathrm{E}(A_x|\mathrm{Sel}_{t-1}) = h_0^2 V_{P_0}^{1/2} i + h_1^2 V_{P_1}^{1/2} i + \cdots + h_{t-1}^2 V_{P_{t-1}}^{1/2} i$$

and the variance of A_x is

$$\mathrm{Var}(A_x|\mathrm{Sel}_{t-1}) = \tfrac{1}{2}V_{A_{t-1}}[1 - h_{t-1}^2 i(i-s)] + \tfrac{1}{2}V_{A_0}$$

4.7.6 Additive Variance at Equilibrium

At equilibrium, $V_{A_{t-1}} = V_{A_t}$. So, if V_{A_e} is the equilibrium variance,

$$V_{A_e} = \tfrac{1}{2}V_{A_e}[1 - h_e^2 i(i-s)] + \tfrac{1}{2}V_{A_0}$$

where

$$h_e^2 = \frac{V_{A_e}}{V_{A_e} + V_D + V_E}$$

Solving for V_{A_e} gives

$$V_{A_e} = \frac{-(V_D + V_E - V_{A_0}) \pm \sqrt{(V_D + V_E - V_{A_0})^2 + 4(1+k)V_{A_0}(V_D + V_E)}}{2(1+k)}$$

where $k = i(i-s)$.

4.7.7 Numerical Example

Assumptions:
$V_A = V_D + V_E = 100$
proportion selected = 0.05
Parents selected by truncation from generation 0-4.

Generation	V_A	E(A)
0	100	0
1	78	14
2	74	26.7
3	74	38.3
4	73	49.8
5	73	61.3
Selection relaxed		
6	87	61.3
7	93	61.3
8	97	61.3
9	98	61.3
10	99	61.3

4.7.8 Genetic Interpretation of Results

There are two contributions to the change in genetic variance by selection:

1. due to change in gene frequency
2. due to covariances between between additive effects within gametes

It can be shown that the contribution to the change in genetic variance due to change in gene frequency goes to zero as the number of loci goes to infinity.
Assume:
n loci with the same allelic effects and frequencies
two alleles a_1 and a_2 with effects α_1 and α_2 at each locus
frequency of a_1 is p and frequency of a_2 is $(1-p)$
Mean of allelic effects before selection:

$$\begin{aligned}\mu_\alpha &= p\alpha_1 + (1-p)\alpha_2 \\ &= 0\end{aligned}$$

Variance of allelic effects before selection:

$$\begin{aligned} V_\alpha &= p(1-p)(\alpha_1 - \alpha_2)^2 \\ &= p(1-p)\alpha^2 \end{aligned}$$

where $\alpha = \alpha_1 - \alpha_2$.
Let the change in gene frequency due selection be denoted by Δ_p. Now, the change in μ_α due to selection is

$$\begin{aligned} \Delta_{\mu_\alpha} &= (p + \Delta_p)\alpha_1 + (1 - p - \Delta_p)\alpha_2 - 0 \\ &= p\alpha_1 + (1-p)\alpha_2 + (\alpha_1 - \alpha_2)\Delta_p \\ &= \alpha\Delta_p \end{aligned} \tag{80}$$

So, Δ_p can be written as

$$\Delta_p = \frac{\Delta_{\mu_\alpha}}{\alpha} \tag{81}$$

Because all n loci have the same allelic effects and frequencies, the change in the mean of A can be written as

$$\Delta_{\mu_A} = 2n\Delta_{\mu_\alpha} \tag{82}$$

So, Δ_{μ_α} can be written as

$$\begin{aligned} \Delta_{\mu_\alpha} &= \frac{\Delta_{\mu_A}}{2n} \\ &= \frac{ih^2 V_P^{1/2}}{2n} \end{aligned} \tag{83}$$

Substituting (83) in (81) gives

$$\Delta_p = \frac{ih^2 V_P^{1/2}}{2n\alpha} \tag{84}$$

Further, because all n loci have the same allelic effects and frequencies, the variance before selection can be written as

$$\begin{aligned} V_A &= 2nV_\alpha \\ &= 2np(1-p)\alpha^2 \end{aligned}$$

So,

$$\alpha = \sqrt{\frac{V_A}{2np(1-p)}} \tag{85}$$

Substituting (85) in (84) gives

$$\Delta_p = ih\sqrt{\frac{p(1-p)}{2n}}$$

So, as $n \to \infty$ $\Delta_p \to 0$.
Finally, the effect of change in gene frequency on the variance is

$$\begin{aligned}\Delta_{V_I} &= 2n(p+\Delta_p)(1-p-\Delta_p)\alpha^2 - 2np(1-p)\alpha^2 \\ &= 2n\alpha^2[\Delta_p(1-p) - p\Delta_p - \Delta_p^2] \\ &= 2n\alpha^2\Delta_p(1-2p-\Delta_p)\end{aligned} \tag{86}$$

Substituting (85) for α in (86) gives

$$\Delta_{V_I} = \frac{V_A\Delta_p(1-2p-\Delta_p)}{p(1-p)}$$

But, as $n \to \infty$, $\Delta_p \to 0$. So, as $n \to \infty$, $\Delta_{V_I} \to 0$.

4.8 Response to Selection in a Finite Population

As we have seen in section 4.7.5, in an infinite population, under normality, response to selection continues indefinitely. In a finite population, however, due to loss of alleles, response to selection is finite. Below, we derive this selection limit for a normally distributed trait that is additively inherited.
Suppose that two alleles are segregating at each locus. To simplify the notation, the difference between the homozygotes at each locus is denoted by 2α. From (80) and (82), the limiting change in the additive genetic mean at some locus is

$$\Delta_{\mu_A} = \sum_j 2\alpha\Delta_{p_\infty}, \tag{87}$$

where the summation is over all loci, and Δ_{p_∞} is the limiting response to selection in gene frequency at locus j. Substituting (73) in (87) gives

$$\Delta_{\mu_A} = \sum_j 2\alpha N_e sp(1-p), \tag{88}$$

where N_e is the effective population size, s is the coefficient of selection for the unfavorable homozygote at locus j, and p is the initial

gene frequency at locus j. From figure 11.6 in the Falconer and Mackay, the coefficient of selection can be approximated by

$$s \approx \frac{i2\alpha}{V_p^{1/2}}. \tag{89}$$

Using this in (88) gives

$$\begin{aligned} \Delta_{\mu_A} &= \sum_j 2\alpha N_e \frac{i2\alpha}{V_p^{1/2}} p(1-p) \\ &= 2N_e i \frac{\sum_j 2\alpha^2 p(1-p)}{V_p^{1/2}} \\ &= 2N_e i \frac{V_A}{V_p^{1/2}}, \end{aligned} \tag{90}$$

which is $2N_e$ times the initial response. This formula shows that the limiting response can be increased by increasing N_e or by increasing i. But, N_e and i are inversely related. It is shown below that the product $N_e i$ is maximum when the top half of the individuals are selected as parents.
If the following, let T be the size of the population, p the proportion of individuals selected to be parents, and N_e the number of parents, which can be written as

$$N_e = Tp.$$

If the trait is normally distributed,

$$i = \frac{z}{p},$$

where z is the ordinate of the standard normal curve at the standardized truncation point s. Thus, $N_e i$ can be written as

$$\begin{aligned} N_e i &= Tpi \\ &= Tz. \end{aligned}$$

In the above, T is a constant, and $N_e i$ can be maximized by maximizing z. The maximum value of z is obtained by selecting the top half of the population to be the parents, and this maximizes the limiting response to selection.

5 Genetic Evaluation

5.1 Minimize Mean Squared Error of Prediction

The genotypic value G is unobservable, and observable phenotypic values $\boldsymbol{y}$ are used to predict G. The predictor $\tilde{G}$ should be some function of $\boldsymbol{y}$, such that

$$\mathrm{E}(G - \tilde{G})^2$$

is minimum. Let

$$\hat{G} = \mathrm{E}(G|\boldsymbol{y}).$$

Now write,

$$\begin{aligned}\mathrm{E}(G - \tilde{G})^2 &= \mathrm{E}(G - \hat{G} + \hat{G} - \tilde{G})^2 \\ &= \mathrm{E}[(G - \hat{G})^2 + (\hat{G} - \tilde{G})^2 \\ &+ \quad 2(G - \hat{G})(\hat{G} - \tilde{G})]\end{aligned}$$

But,

$$\begin{aligned}\mathrm{E}(G - \hat{G})(\hat{G} - \tilde{G}) &= \underset{\boldsymbol{y}}{\mathrm{E}}[\mathrm{E}(G - \hat{G})(\hat{G} - \tilde{G})|\boldsymbol{y}] \\ &= \underset{\boldsymbol{y}}{\mathrm{E}}\{(\hat{G} - \tilde{G})\mathrm{E}[(G - \hat{G})|\boldsymbol{y}]\} \\ &= \underset{\boldsymbol{y}}{\mathrm{E}}[(\hat{G} - \tilde{G})(\hat{G} - \hat{G})] \\ &= 0\end{aligned}$$

Then,

$$\mathrm{E}(G - \tilde{G})^2 = \mathrm{E}(G - \hat{G})^2 + \mathrm{E}(\hat{G} - \tilde{G})^2$$

The first term does not depend on $\tilde{G}$
The second term is minimized by choosing

$$\tilde{G} = \hat{G}$$

So, $\hat{G}$ is the best predictor of G

5.2 Conditional Mean Under Normality

Consider a vector $\boldsymbol{y}$ with three phenotypic values. Can show that under normality,

$$\hat{G} = \mu + b_1(y_1 - \mu) + b_2(y_2 - \mu) + b_3(y_3 - \mu)$$

The b_i are obtained by solving:

$$\begin{array}{ccccccc} b_1V_{11} & + & b_2V_{12} & + & b_3V_{13} = C_1 \\ b_1V_{21} & + & b_2V_{22} & + & b_3V_{23} = C_2 \\ b_1V_{31} & + & b_2V_{32} & + & b_3V_{33} = C_3 \end{array}$$

where $V_{ij} = \text{Cov}(y_i, y_j)$ and $C_i = \text{Cov}(y_i, G)$.

$$G = \hat{G} + \epsilon$$

where $\epsilon = (G - \hat{G})$. Observe that

$$\text{E}(\epsilon) = 0$$

$$\begin{aligned} \text{Cov}[\epsilon, y_i] &= C_i - (b_1V_{i1} + b_2V_{i2} + b_3V_{i3}) \\ &= 0 \end{aligned}$$

Thus, under normality,

$$\begin{aligned} \text{E}(G|\boldsymbol{y}) &= \text{E}(\hat{G}|\boldsymbol{y}) + \text{E}(\epsilon|\boldsymbol{y}) \\ &= \hat{G} + \text{E}(\epsilon) \\ &= \hat{G} \end{aligned}$$

5.3 Maximize Correlation between G and $\tilde{G}$

It is shown below that

$$\rho(G, \tilde{G}) = \frac{\text{Cov}(G, \tilde{G})}{\sqrt{\text{Var}(G)\text{Var}(\tilde{G})}} \tag{91}$$

is maximized by choosing $\tilde{G} = \widehat{G}$. Let $\text{E}(\tilde{G}) = \theta$. Then,

$$\begin{aligned} \text{Cov}(G, \tilde{G}) &= \text{E}\left[G(\tilde{G} - \theta)\right] \\ &= \text{E}\left\{\left[(G - \widehat{G}) + \widehat{G}\right](\tilde{G} - \theta)\right\} \end{aligned} \tag{92}$$

But,

$$\begin{aligned}\mathrm{E}_{\boldsymbol{y}}\left\{\mathrm{E}\left[(G-\widehat{G})(\tilde{G}-\theta)\mid \boldsymbol{y}\right]\right\} &= \mathrm{E}_{\boldsymbol{y}}\left\{(\tilde{G}-\theta)\mathrm{E}\left[(G-\widehat{G})\mid \boldsymbol{y}\right]\right\} \\ &= \mathrm{E}_{\boldsymbol{y}}\left[(\tilde{G}-\theta)(\widehat{G}-\widehat{G})\right]=0\end{aligned} \tag{93}$$

So

$$\mathrm{Cov}(G,\tilde{G}) = \mathrm{E}\left[\widehat{G}(\tilde{G}-\theta)\right] = \mathrm{Cov}(\widehat{G},\tilde{G}) \tag{94}$$

Also, $\mathrm{Cov}(G,\widehat{G}) = \mathrm{Cov}(\widehat{G},\widehat{G}) = \mathrm{Var}(\widehat{G})$. Now,

$$\begin{aligned}\rho^2(G,\tilde{G}) &= \frac{\mathrm{Cov}^2(G,\tilde{G})}{\mathrm{Var}(G)\mathrm{Var}(\tilde{G})} \\ &= \frac{\mathrm{Cov}^2(\widehat{G},\tilde{G})}{\mathrm{Var}(G)\mathrm{Var}(\tilde{G})} \\ &= \frac{\mathrm{Cov}^2(\widehat{G},\tilde{G})}{\mathrm{Var}(\widehat{G})\mathrm{Var}(\tilde{G})}\frac{\mathrm{Var}(\widehat{G})}{\mathrm{Var}(G)} \\ &= \rho^2(\widehat{G},\tilde{G})\frac{\mathrm{Var}(\widehat{G})}{\mathrm{Var}(G)}\end{aligned} \tag{95}$$

This is maximum when $\tilde{G}=\widehat{G}$ and $\rho^2(\widehat{G},\tilde{G})=1$. Note that

$$\frac{\mathrm{Var}(\widehat{G})}{\mathrm{Var}(G)} = \rho^2(G,\widehat{G})$$

5.4 Maximize Mean of Selected Candidates

Consider now the problem of maximizing the expected value of selected $G_i's$. Suppose there are n candidates and we want to choose k such that

$$\mathrm{E}\left[\frac{\sum_{i=1}^{k} G_{s_i}}{k}\right]$$

where $s_1,\ldots,s_k$ are the indices of the selected $G_i's$.

$$\mathrm{E}\left[\frac{\sum_{i=1}^{k} G_{s_i}}{k}\right] = \frac{1}{k}\mathrm{E}_{\boldsymbol{y}}\left[\mathrm{E}\left(\sum_{i=1}^{k} G_{s_i}\mid \boldsymbol{y}\right)\right] = \frac{1}{k}\mathrm{E}_{\boldsymbol{y}}\left[\sum_{i=1}^{k}\widehat{G}_{s_i}\right] \tag{96}$$

It is clear that selecting $s_1, \ldots, s_k$ to be the indices of highest ranking $\widehat{G}_i$ would maximize 96. This result is very general. It does not depend on the joint distribution of $\boldsymbol{G}$ and $\boldsymbol{y}$. Here, the proportion selected is a constant.

5.5 Accuracy of Prediction

Accuracy of prediction is given by:

$$\text{Cor}(G, \hat{G}) = \frac{\text{Cov}(G, \hat{G})}{\sqrt{\text{Var}(G)\text{Var}(\hat{G})}},$$

where

$$\begin{aligned}\text{Cov}(G, \hat{G}) &= \text{Cov}(\hat{G} + \epsilon, \hat{G}) \\ &= \text{Cov}(\hat{G}, \hat{G}) \\ &= \text{Var}(\hat{G}).\end{aligned}$$

So,

$$\begin{aligned}\text{Cor}(G, \hat{G}) &= \frac{\text{Var}(\hat{G})}{\sqrt{\text{Var}(G)\text{Var}(\hat{G})}} \\ &= \sqrt{\frac{\text{Var}(\hat{G})}{\text{Var}(G)}}.\end{aligned}$$

Under normality,

$$\hat{G} = \mu + b_1(y_1 - \mu) + b_2(y_2 - \mu) + b_3(y_3 - \mu),$$

$$\begin{aligned}\text{Var}(\hat{G}) &= b_1^2 V_{11} + b_1 b_2 V_{12} + b_1 b_3 V_{13} \\ &\quad + b_2 b_1 V_{21} + b_2^2 V_{22} + b_2 b_3 V_{13} \\ &\quad + b_3 b_1 V_{31} + b_3 b_2 V_{32} + b_3^2 V_{33} \\ &= b_1 C_1 + b_2 C_2 + b_3 C_3\end{aligned}$$

5.6 Example

Example 17

Consider an additive trait with $V_A = 1$ and $V_P = 4$
Want to predict G_x given P_x, P_s, and P_d
The index equations are:

$$\begin{array}{ccccccc} b_1 4.0 & + & b_2 0.5 & + & b_3 0.5 & = & 1.0 \\ b_1 0.5 & + & b_2 4.0 & + & b_3 0.0 & = & 0.5 \\ b_1 0.5 & + & b_2 0.0 & + & b_3 4.0 & = & 0.5 \end{array}$$

The solution is: $b_1 = 0.2258$, $b_2 = b_3 = 0.0968$
$\mathrm{Var}(\hat{G}) = 0.3226$ and $\mathrm{Cor}(G, \hat{G}) = 0.5680$

Example 18

Suppose in addition to P_x, P_s, and P_d, the mean ($\bar{P}_o$) of n offspring records are available.
The covariance of $\bar{P}_o$ with the other phenotypic records does not depend on n. For example:

$$\begin{aligned} \mathrm{Cov}(P_x, \bar{P}_o) &= \mathrm{Cov}(P_x, \frac{P_{o_1} + P_{o_2} + \cdots + P_{o_n}}{n}) \\ &= \frac{0.5V_A + 0.5V_A + \cdots + 0.5V_A}{n} \\ &= 0.5V_A \end{aligned}$$

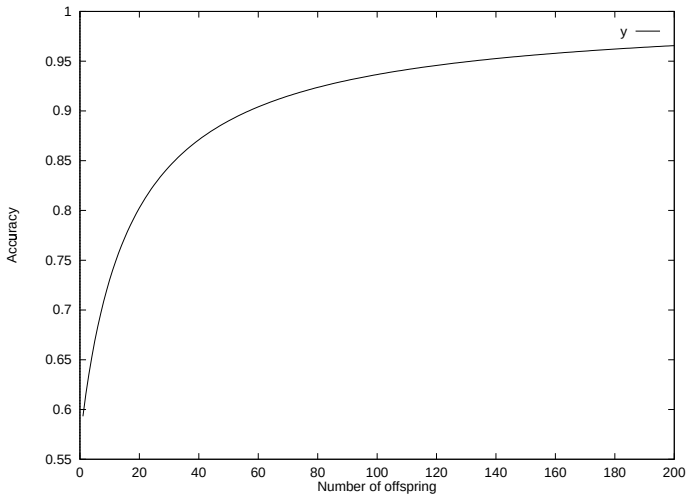
The variance of $\bar{P}_o$, however, depends on n.
Suppose the n offspring are half-sibs. Then:

$$\begin{aligned} \mathrm{Var}(\bar{P}_o) &= \mathrm{Var}(\frac{P_{o_1} + P_{o_2} + \cdots + P_{o_n}}{n}) \\ &= \frac{1}{n^2}(n4.0 + n(n-1)0.25) \\ &= \frac{4.0 + (n-1)0.25}{n} \end{aligned}$$

The index equations are:

$$\begin{array}{ccccccccc} b_1 4.0 & + & b_2 0.5 & + & b_3 0.5 & + & b_4 0.5 & = & 1.0 \\ b_1 0.5 & + & b_2 4.0 & + & b_3 0.0 & + & b_4 0.25 & = & 0.5 \\ b_1 0.5 & + & b_2 0.0 & + & b_3 4.0 & + & b_4 0.25 & = & 0.5 \\ b_1 0.5 & + & b_2 0.25 & + & b_3 0.25 & + & b_4 \mathrm{Var}(\bar{P}_o) & = & 0.5 \end{array}$$

Figure 1: Accuracy of prediction with n offspring (halfsib) records in addition to records on parents and self.



6 Estimation of Genetic Parameters

7 Inbreeding Depression and Heterosis

8 QTL Mapping

8.1 QTL Mapping Using Line Crosses

8.1.1 Difference between marker groups

Consider two inbred lines P_1 and P_2 where all individuals in P_1 have genotype $\frac{Q_1A_1}{Q_1A_1}$ and all individuals in P_2 have genotype $\frac{Q_2A_2}{Q_2A_2}$, for QTL Q and marker locus A. When these two lines are crossed, all the F_1 individuals will have genotype $\frac{Q_1A_1}{Q_2A_2}$.
To investigate if A and Q are linked, we obtain trait (y) and marker data from either the F_2 or a backcross generation. Suppose data are obtained from the backcross BC_1 obtained by crossing F_1 with P_1. Individuals in BC_1 will get the gamete Q_1A_1 with probability 1 from the P_1 parent, and they will get Q_1A_1 with probability $\frac{1}{2}(1-r)$,

Q_1A_2 with probability $r/2$, Q_2A_1 with probability $r/2$, or Q_2A_2 with probability $\frac{1}{2}(1-r)$, from the F_1 parent. The BC_1 individuals can be divided into two groups: those with marker genotype $\frac{A_1}{A_1}$ and those with genotype $\frac{A_1}{A_2}$. A BC_1 individual with genotype $\frac{A_1}{A_1}$ will have QTL genotype $\frac{Q_1}{Q_1}$ with probability $(1-r)$ or genotype $\frac{Q_1}{Q_2}$ with probability r. Similarly, a BC_1 individual with genotype $\frac{A_1}{A_2}$ will have QTL genotype $\frac{Q_1}{Q_1}$ with probability r or genotype $\frac{Q_1}{Q_2}$ with probability $(1-r)$. Thus, the expected value of the trait for a BC_1 individual with genotype $\frac{A_1}{A_1}$ is

$$\mu_{A_1A_1} = (1-r)\mu_{11} + r\mu_{12}$$

where μ_{11} is the expected trait value for an individual with QTL genotype $\frac{Q_1}{Q_1}$ and μ_{12} is the expected trait value for an individual with QTL genotype $\frac{Q_1}{Q_2}$. The expected value of the trait for a BC_1individual with genotype $\frac{A_1}{A_2}$ is

$$\mu_{A_1A_2} = r\mu_{11} + (1-r)\mu_{12}$$

The difference between these expected values is

$$\mu_{A_1A_1} - \mu_{A_1A_2} = (\mu_{11} - \mu_{12})(1-2r) = \delta(1-2r) \tag{97}$$

If the QTL is not linked to the marker, $r = \frac{1}{2}$ and $\mu_{A_1A_1} - \mu_{A_1A_2} = 0$. So, a t-test can be used to test the hypothesis: H_0 $\mu_{A_1A_1} - \mu_{A_1A_2} = 0$ vs. H_a $\mu_{A_1A_1} - \mu_{A_1A_2} \neq 0$. This test will be approximate because the trait has a mixture distribution. Further, there is more than one value of δ and r that will result in the the same value for $\mu_{A_1A_1} - \mu_{A_1A_2}$. Thus, with this analysis, δ and r are confounded.

To calculate the power of this test, we assume that given QTL genotype $\frac{Q_1}{Q_1}$, $y \sim N(\mu_{11}, \sigma^2)$, and given QTL genotype $\frac{Q_1}{Q_2}$, $y \sim N(\mu_{12}, \sigma^2)$. Let y_{1j} be the trait value of individual j with marker genotype $\frac{A_1}{A_1}$ and y_{2j} the trait value of individual j with marker genotype $\frac{A_1}{A_2}$. Then, variance of y_{1j} can be written as

$$\begin{aligned} \mathrm{Var}(y_{1j}) = {} & \mathrm{E}[\mathrm{Var}(y_{1j} \mid \text{QTL genotype})] \\ & + \mathrm{Var}[\mathrm{E}(y_{1j} \mid \text{QTL genotype})] \end{aligned} \tag{98}$$

The first term of (98) is

$$\mathrm{E}[\mathrm{Var}(y_{1j} \mid \text{QTL genotype}] = \sigma^2(1-r) + \sigma^2 r = \sigma^2$$

and the second term of (98) is

$$\text{Var}[\text{E}(y_{1j} \mid \text{QTL genotype}] = (\mu_{11} - \mu_{12})^2(1-r)r = \delta^2(1-r)r$$

So, variance of y_{1j} is

$$\text{Var}(y_{1j}) = \sigma^2 + \delta^2(1-r)r \tag{99}$$

Similarly, it can be shown that variance of y_{2j} is identical to (99). Now, the distribution of the difference between the class means can be approximated as

$$\bar{y}_{1.} - \bar{y}_{2.} \sim N(\delta(1-2r), 2[\sigma^2 + \delta^2(1-r)r]/n)$$

where n is the number of individuals in each marker class. If n is large, the t distribution approaches a normal distribution. So, power will be computed for a normal test. For a normal test the test statistic is

$$Z = \frac{(\bar{y}_{1.} - \bar{y}_{2.})}{\sqrt{2[\sigma^2 + \delta^2(1-r)r]/n}}$$

Now, the power of the test is

$$\Pr(Z > Z_{\alpha/2})$$

where Z_α is the point for which $\Pr(Z > Z_\alpha) = \alpha$. Under H_a the expected value of Z is

$$\text{E}(Z) = \frac{\delta(1-2r)}{\sqrt{2[\sigma^2 + \delta^2(1-r)r]/n}} \tag{100}$$

Subtracting $\text{E}(Z)$ from Z and $Z_{\alpha/2}$ gives

$$\text{Power} = \Pr[Z - \text{E}(Z) > Z_{\alpha/2} - \text{E}(Z)]$$

Note that $Z - \text{E}(Z)$ has a standard normal distribution. Thus, for the power to be $1-\beta$, $Z_{\alpha/2} - \text{E}(Z) = Z_{1-\beta} = -Z_\beta$. Substituting (100) in this expression for $\text{E}(Z)$ and solving for n gives the required sample size for the power to be $1-\beta$:

$$n = (Z_{\alpha/2} + Z_\beta)^2 \frac{2[\sigma^2 + \delta^2(1-r)r]}{\delta^2(1-2r)^2} \tag{101}$$

Example: Consider an additive trait with additive variance σ_a^2 in the F_2 generation. Let p be the proportion of the additive variance in the F_2 due to the QTL. Thus,

$$p = \tfrac{1}{2}\delta^2/\sigma_a^2$$

and

$$\sigma_a^2 = \tfrac{1}{2}\delta^2/p$$

Let h^2 be the heritability in the F_2 generation defined as

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

where $\sigma_e^2 = \sigma_a^2(1-h^2)/h^2$ is the environmental variance. The additive variance in the backcross is half that in the F_2. So, the total variance in the backcross generation, given the genotype at the QTL is

$$\begin{aligned} \sigma^2 &= \sigma_e^2 + \tfrac{1}{2}(\sigma_a^2 - \tfrac{1}{2}\delta^2) \\ &= \tfrac{1}{2}\delta^2\frac{(1-h^2)}{ph^2} + \frac{1}{4}\delta^2(\frac{1}{p} - 1) \\ &= \delta^2[\tfrac{1}{2}\frac{(1-h^2)}{ph^2} + \frac{1}{4}(\frac{1}{p} - 1)] \end{aligned} \tag{102}$$

Substituting (102) for σ^2 in (101) gives

$$n = 2\frac{(Z_{\alpha/2} + Z_\beta)^2}{(1-2r)^2}\{\tfrac{1}{2}\frac{(1-h^2)}{ph^2} + \frac{1}{4}(\frac{1}{p} - 1) + r(1-r)\}$$

For a trait with $h^2 = 0.25$, the sample sizes required for power of test to be 0.9 are given in table (1).

8.1.2 Regression

As we have seen from (97) the difference between marker genotype classes cannot be used to estimate the recombination rate or the QTL effects. These parameters can be estimated by a regression method that will be outlined here. The method will be described for use with backcross data.

Table 1: Sample size ($2n$) required for power of test to be 0.90 in a backcross experiment to detect a QTL that contributes a proportion p to additive genetic variance. The recombination rate between the QTL and marker is r and the significance level is $\alpha = 0.05$

	p		
r	0.04	0.08	0.16
0	1,828	909	449
0.05	2,260	1,125	557
0.1	2,863	1,426	708
0.2	5,097	2,543	1,266
0.4	45,959	22,974	11,482

Assumptions and notation: Individuals in inbred line P_1 have genotype $\frac{A_1Q_1B_1}{A_1Q_1B_1}$, and those in line P_2 have genotype $\frac{A_2Q_2B_2}{A_2Q_2B_2}$. Recombination fraction between marker locus A and QTL Q is r_A, between Q and marker locus B is r_B, and between A and B is r_{AB}; r_{AB} is assumed to be known. In the backcross generation, the expected value of the phenotypic value (y), given QTL genotypes are

$$\mathrm{E}(y \mid Q_1Q_1) = \mu_1$$

$$\mathrm{E}(y \mid Q_1Q_2) = \mu_2$$

The variance is assumed to be the same in both QTL genotype classes.

Theory: In the F_1 all individuals have genotype $\frac{A_1Q_1B_1}{A_2Q_2B_2}$. Individuals in backcross $B1$ produced by mating F_1 with P_1 will have four marker genotypes and two QTL genotypes. Assuming the Haldane mapping function, conditional probabilities for these QTL genotypes given the marker genotypes are given in table (2).
Now, the expected value of the trait phenotypic values can be written as

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of phenotypic values, $\boldsymbol{X}$ is an $n \times 2$ matrix of probabilities from table (2), and $\boldsymbol{\beta}$ has the unknown genotypic means: μ_1, and μ_2. If individual i has marker genotype j, the ith row of $\boldsymbol{X}$ will contain the probabilities from the jth row of table

Table 2: Conditional probabilities of the QTL genotypes given marker genotypes in backcross generation.

Marker Genotype	QTL Genotype	
	$\frac{Q_1}{Q_1}$	$\frac{Q_1}{Q_2}$
$\frac{A_1B_1}{A_1B_1}$	$\frac{(1-r_A)(1-r_B)}{(1-r_{AB})}$	$\frac{r_A r_B}{(1-r_{AB})}$
$\frac{A_1B_1}{A_1B_2}$	$\frac{(1-r_A)r_B}{r_{AB}}$	$\frac{r_A(1-r_B)}{r_{AB}}$
$\frac{A_1B_1}{A_2B_1}$	$\frac{r_A(1-r_B)}{r_{AB}}$	$\frac{(1-r_A)r_B}{r_{AB}}$
$\frac{A_1B_1}{A_2B_2}$	$\frac{r_A r_B}{(1-r_{AB})}$	$\frac{(1-r_A)(1-r_B)}{(1-r_{AB})}$

(2). Given the QTL position, $\boldsymbol{X}$ can be computed and $\boldsymbol{\beta}$ estimated by least squares as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

The residual sum of squares is given by

$$\text{RSS} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

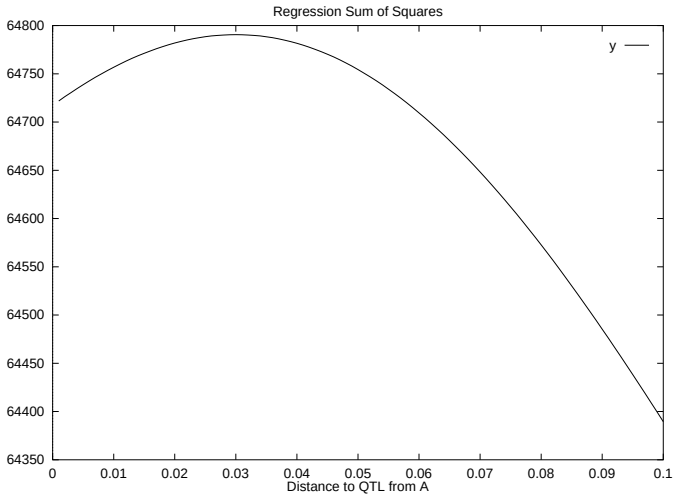
The position that minimizes the residual sum of squares gives the estimated position of the QTL.

Example: Regression sums of squares were computed for 100 evenly spaced locations of a QTL between two markers 10 cM apart (Figure 2). The trait means and sample sizes for the four marker genotype classes were set to their expected values for $\mu_1 = 20, \mu_2 = 30$ and a map distance of 3 cM from marker A to the QTL.

9 QTL Mapping in Outbred Populations

We have already seen that even loci that are linked can be in equilibrium due to random mating. When a marker is in equilibrium with the QTL, the conditional distribution of the QTL given the marker is identical to its unconditional distribution. Thus, methods used for mapping QTL with line cross data, which rely on marker genotype classes having different means, are not suitable for mapping QTL in outbred populations, unless the data are from a single family.

Figure 2: Plot of regression sum of squares corresponding to position of QTL between two markers 10 cM apart.



Although a marker that is in equilibrium with the QTL cannot be used to model means, the observed marker genotypes of relatives can be used to model genetic covariances between them. For example, if two halfsibs receive the same marker allele from their common parent, then both sibs are likely to receive the same allele from a QTL that is closely linked to this marker. This causes them to be more highly correlated than two sibs that received different marker alleles. This relationship between the observed marker genotypes and genetic covariances has been used to develop methods for mapping QTL in outbred populations.

More generally, even when the marker and QTL are in equilibrium, the joint distribution of QTL genotypes of relatives depends on the observed marker data. Thus, maximum likelihood methods can be used to map QTL in outbred populations. Some of these methods are described below.

9.1 Halfsib data with one marker

Consider halfsib data from unrelated sires that are heterozygous for a marker locus A. Suppose a sire has genotype $A_jA_{j'}$. If an offspring from this sire can be classified as receiving allele A_j or allele $A_{j'}$ from the sire, it is said to be informative. The following model will be used for analyzing data from informative offspring:

$$y_{ijk} = \mu + s_i + m_{ij} + e_{ijk} \tag{103}$$

where y_{ijk} is the trait phenotype of offspring k that received marker j from sire i, $\mu = \mathrm{E}(y_{ijk})$, s_i is a random effect with null mean and variance σ^2_s, m_{ij} is a random effect with null mean and variance σ^2_m, and e_{ijk} is a random residual with null mean and variance σ^2_e. The random effects are assumed to be independent.
Given this model, it follows that the covariance between two halfsibs that received the same marker allele from their sire would be

$$\mathrm{Cov}(y_{ijk}, y_{ijk'}) = \sigma^2_s + \sigma^2_m \tag{104}$$

Suppose marker A is linked to a QTL that contributes σ^2_Q to the additive genetic variance. Then the above covariance (104) can also be written as

$$\mathrm{Cov}(y_{ijk}, y_{ijk'}) = [(1-r)^2 + r^2]\sigma^2_Q/2 + \sigma^2_u/4 \tag{105}$$

where r is the probability of recombination between the marker locus and the QTL. Further, from (103), the covariance between two halfsibs that received different marker alleles from their sire would be

$$\mathrm{Cov}(y_{ijk}, y_{ij'k'}) = \sigma^2_s \tag{106}$$

and assuming A is linked to the QTL, this covariance (106) can be written as

$$\mathrm{Cov}(y_{ijk}, y_{ij'k'}) = 2r(1-r)\sigma^2_Q/2 + \sigma^2_u/4 \tag{107}$$

From (104), (105), (106) and (107), it follows that

$$\begin{aligned}\sigma^2_m &= [(1-r)^2 + r^2 - 2r(1-r)]\sigma^2_Q/2 \\ &= (1-2r)^2\sigma^2_Q/2\end{aligned}$$

and σ_m^2 will not be null unless $r = \frac{1}{2}$. To test the if $\sigma_m^2 > 0$, we compute

$$F_{\text{cal}} = \frac{\text{MS}_m}{\text{MS}_e} \tag{108}$$

where MS_m is the mean squares for marker within sire and MS_e the mean square for error. Under the H_0: $\sigma_m^2 = 0$, F_{cal} has a central F_{ν_1,ν_2} distribution, where $\nu_1 = n_s$, $\nu_2 = 2n_s \sum(n_{ij} - 1)$, n_s is the number of sires, and n_{ij} is the number of offspring that received marker j from sire i. Under the alternative hypothesis, F_{cal} is distributed as

$$F_{\text{cal}} \sim F_{\nu_1,\nu_2} \frac{\text{E}(\text{MS}_m)}{\text{E}(\text{MS}_e)}$$

Suppose $n_{ij} = n_w$, then $\text{E}(\text{MS}_m) = \sigma_e^2 + n_w\sigma_m^2$, and $\text{E}(\text{MS}_e) = \sigma_e^2$. Thus, the power of the test is

$$\Pr(F_{\nu_1,\nu_2} > F_{\alpha,\nu_1,\nu_2} \frac{\sigma_e^2}{\sigma_e^2 + n_w\sigma_m^2})$$

where F_{α,ν_1,ν_2} is the value for which $\Pr(F_{\nu_1,\nu_2} > F_{\alpha,\nu_1,\nu_2}) = \alpha$.

10 Marker Assisted Selection

11 Appendix

11.1 Binomial Distribution

Consider a random variable X with:

$$\Pr(X = 1) = q,$$

and

$$\Pr(X = 0) = 1 - q.$$

This is called a Bernoulli random variable. The expected value of X is

$$\begin{aligned} \mathrm{E}(X) &= 0(1-q) + 1q \\ &= q. \end{aligned} \tag{109}$$

The variance of X is

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2,$$

where

$$\begin{aligned} \mathrm{E}(X^2) &= 0^2(1-q) + 1^2 q \\ &= q \end{aligned}$$

So, the variance of X is

$$\begin{aligned} \mathrm{Var}(X) &= q - q^2 \\ &= q(1-q) \end{aligned} \tag{110}$$

Now let

$$Y = \sum_{i=1}^{n} X_i,$$

where X_i are identically and independently distributed Bernoulli random variables. Then, Y is said to have a Binomial distribution with parameters n and q, and denoted

$$Y \sim \mathrm{Binomial}(n, q)$$

The expected value of Y is

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(X_1 + X_2 + \cdots + X_n) \\ &= q + q + \cdots + q \\ &= nq, \end{aligned} \tag{111}$$

and the variance of Y is

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots \mathrm{Var}(X_n) \\ &= nq(1-q) \end{aligned} \tag{112}$$

The probability distribution for a Binomial random variable is

$$\Pr(Y = y) = \frac{n!}{(n-y)!y!} q^y (1-q)^{n-y} \tag{113}$$

Example 19 *Consider a population where the frequency of allele A_2 is 0.2. Suppose 20 gametes are sampled from this population. When gamete i is sampled, put $X_i = 1$ if the allele at locus A is A_2 and put $X_i = 0$ if it is A_1. Then,*

$$Y = \sum_{i=1}^{20} X_i$$

is the number of A_2 alleles sampled. Further, $Y \sim$ Binomial$(20, 0.2)$. So,

$$\begin{aligned} E(Y) &= 20 \times 0.2 \\ &= 4, \end{aligned} \tag{114}$$

and

$$\begin{aligned} Var(Y) &= 20 \times 0.2(1 - 0.2) \\ &= 3.2 \end{aligned} \tag{115}$$

11.2 Geometric Series

Let S_n be the geometric series:

$$S_n = 1 + x + x^2 + x^3 + \cdots + x^{n-1} \tag{116}$$

Then, xS_n is

$$xS_n = x + x^2 + x^3 + \cdots + x^{n-1} + x^n \tag{117}$$

Subtracting (117) from (116) gives

$$\begin{aligned} S_n(1-x) &= 1 - x^n \\ S_n &= \frac{1-x^n}{1-x} \end{aligned} \tag{118}$$