Chapter 1

Quantitative 1

In the chapter on Mendel and Morgan, we saw how the transmission of genes from one generation to another follows a precise mathematical formula. The traits discussed in that chapter, however, were discrete traits—peas are either yellow or green, someone either has a disorder or does not have a disorder. But many behavioral traits are not like these clear-cut, have-it-or-don’t-have-it phenotypes. People vary from being quite shy to very outgoing. But is shyness a discrete trait or merely a descriptive adjective for one end of a continuous distribution? In this chapter, we will discuss the genetics of quantitative, continuously distributed phenotypes.

Let us note first that genetics has made important—albeit not widely recognized—contributions to quantitative methodology in the social sciences. The concept of regression was initially developed by Sir Francis Galton in his attempt to predict offspring phenotypes from parental phenotypes; it was later expanded and systematized by his colleague, Karl Pearson, in the context of evolutionary theory. The analysis of variance was formulated by Sir Ronald A. Fisher to solve genetic problems in agriculture. Finally, the famous American geneticist Sewell Wright developed the technique of path analysis, which is now used widely in psychology, sociology, anthropology, economics, and other social sciences.

1.1 Genetic Variance Components: Introduction

In the discussion of variance in Chapter X.X, we noted that variance is “statistical” for individual differences and the concept is important because it can be partitioned. Think of variance as being a “pie” of individual differences. We want to partition the pie based on data into parts due to different genetic effects and different environmental effects. These parts are called variance components.
1.1.1 Single locus model

Let us begin the development of a quantitative model by considering a single gene with two alleles, \( a \) and \( A \). Define the genotypic value (aka genetic value) for a genotype as the average phenotypic value for all individuals with that genotype. For example, suppose that the phenotype was IQ, and we measured IQ on a very large number of individuals. Suppose that we also genotyped these individuals for the locus. The genotypic value for genotype \( aa \) would be the average IQ of all individuals who had genotype \( aa \). Hence, the means for genotypes \( aa, Aa, \) and \( AA \) would be different from one another, but there would still be variation around each genotype. This situation is depicted in Figure 1.1.

Figure 1.1: A single locus example for IQ.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
<td>70</td>
<td>85</td>
<td>100</td>
<td>130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first point to notice about Figure 1.1 is the variation in IQ around each of the three genotypes, \( aa, Aa, \) and \( AA \). Not everyone with genotype \( aa \), for instance, has the same IQ. The reasons for this variation within each genotype are unknown. They would include environmental variation as well as the effects of loci other than the one genotyped.

A second important feature about Figure 1.1 is that the means of the distributions for the three genotypes differ. The mean IQ (i.e., the genotypic value) for \( aa \) is 94, that for \( Aa \) is 96, and the mean for \( AA \) is 108. This implies that the locus has some influence on individual differences in IQ.

A third feature of note in Figure 1.1 is that the genotypic value of heterozygote is not equal to the average of the genotypic values of the two homozygotes. The average value of genotypes \( aa \) and \( AA \) is \((94 + 108)/2 = 101\), but the actual genotypic value of \( Aa \) is 96. This indicates a certain degree of dominant gene action for allele \( a \). Allele \( a \) is not completely dominant; otherwise, the genotype value for \( Aa \) would equal that of \( aa \). Hence, the degree of dominance is incomplete.

A fourth feature of importance is that the curves for the three genotypes do not achieve the same height. This is due to the fact that the three genotypes have different frequencies. In the calculations used to generate the figure, it was assumed that the allele frequency for \( a \) was .4 and the frequency for \( A \) was .6, giving the genotypic frequencies as .16 (\( aa \)), .48 (\( Aa \)), and .36 (\( AA \)). Consequently, the curve for \( Aa \) has the highest peak, the one for \( AA \) has the second highest peak, and that for \( aa \) has the smallest peak.

A final feature of note is that the phenotypic distribution of IQ in the general population (the solid line in Figure 1.1) looks very much like a normal distribution. The phenotypic distribution is simply the sum of the distributions for the three genotypes. For example, the height of the curve labeled “Total” when IQ
CHAPTER 1. QUANTITATIVE VARIANCE COMPONENTS: INTRODUCTION

Table 1.1: Schema for genetic values of two loci influencing IQ.

<table>
<thead>
<tr>
<th>Locus A:</th>
<th>Locus B:</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>bb</td>
<td>101</td>
<td>106</td>
<td>111</td>
</tr>
<tr>
<td>Aa</td>
<td>Bb</td>
<td>89</td>
<td>94</td>
<td>99</td>
</tr>
<tr>
<td>aa</td>
<td>BB</td>
<td>87</td>
<td>92</td>
<td>97</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>93</td>
<td>98</td>
<td>103</td>
</tr>
</tbody>
</table>

equals 90 is the distance from the horizontal axis at 90 to the curve for genotype aa plus the distance from the horizontal axis at 90 to the curve for genotype Aa plus the distance from the horizontal axis at 90 to the curve for genotype AA. Often social scientists mistakenly conclude that the phenotypic distribution must be trimodal because it is the sum of three different distributions.

The gene depicted in Figure 1.1 is currently termed a QTL for Quantitative Trait Locus. Behavioral genetic research devotes considerable effort toward uncovering QTLs for many different traits—intelligence, reading disability, various personality traits, and psychopathology. The mathematical models that quantify the extent to which a QTL contributes to trait variance are not necessary for us to know.

Finally, there do not appear to be any genes for IQ that have an effect as large as the one depicted in the Figure. This example is for illustrative purposes only.

1.1.2 Multiple Loci

Genetic individual differences in IQ are due to many genes. Let us begin examination of the effects of multiple loci by considering a second locus, say the B locus with its two alleles, b and B. We would now have nine genotypic values—the three genotypes at locus A combined with the three at locus B—giving the three by three table illustrated in Table 1.1. Once again, we would compute the mean IQ score for all those with a genotype of aabb and then enter this mean into the appropriate cell of the table. Next we could compute the genetic value for all those with genotype Aabb and enter this value into the table and so on. The results would—hypothetically at least—be similar to the data given in the table.

We could also draw curves for each genotype analogous to the curves depicted in Figure 1.1.1. This time, however, there would be nine normal curves, one for each genotype.

We could continue by adding a third locus with two alleles. This would give 27 different genotypes and 27 curves. If we could identify each and every locus that contributes to IQ, then we would probably have a very large number of curves. If we plotted all of the genotypes, then the variation within each curve would be due to the environment.

This model is equivalent to the polygenic model introduced in Chapter X.X.
1.2 Genetic Variance Components: Estimation (Graduate)

1.2.1 The quantitative model (graduate)

The simplest quantitative model involves a single locus with two alleles which we shall designate as $a$ and $A$. The model begins by calculating the phenotypic mean of the three genotypes, $aa$, $Aa$, and $AA$, which for the sake of example, we will assume have the respective values of 1, 1.7, and 2. We express these mean in terms of the algebraic quantities given in Figure 1.2. Here, $m$ is the midpoint between the two homozygotes. For our example, $m = (2 - 1)/2 = 1.5$.

The quantity $\alpha$ in Figure 1.2 is the displacement of the two homozygotes from the midpoint. Given that $m + \alpha = 2$ and $m = 1.5$, then $\alpha = 0.5$. The quantity $\delta$ is the displacement of the heterozygote mean from the midpoint. Because $m + \delta = 1.7$, $\delta = 0.7$. The mean phenotypes for each genotype as well as the algebraic quantities that equal them in Figure 1.2 are called the genetic or genotypic values for the locus.

To calculate the mean and variance due to this locus, we must also consider the frequencies of the two alleles. Let the quantity $p$ equal the frequency of allele $A$ and $q = (1 - p)$, the frequency of allele $a$. Then, if the genotypes consist of the random pairing of alleles, the frequency of genotypes $aa$, $Aa$, and $AA$ will be respectively $q^2$, $2pq$, and $p^2$. The overall phenotypic mean can be expressed in algebra by multiplying the genotypic frequency by the mean phenotypic value for a genotypic and then summing over phenotypes, or

$$
\mu = q^2(m - \alpha) + 2pq(m + \delta) + p^2(m + \alpha)
$$

(1.1)

$$
= m + (p - q)\alpha + 2pq\delta
$$

since $q^2 + 2pq + p^2 = 1.0$ and $p^2 - q^2 = p - q$.

To calculate the variance due to this locus, we must first derive the deviation of each genetic value from the mean. For genotype $aa$ this equals

$$
m - \alpha - \mu =
$$
Table 1.2: Notation for calculations at a single locus.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Freq</th>
<th>Value</th>
<th>Deviation from the mean</th>
<th>Squared deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>(q^2)</td>
<td>(m - a)</td>
<td>(-2p(a - q\delta))</td>
<td>(4p^2(a - q\delta)^2)</td>
</tr>
<tr>
<td>Aa</td>
<td>(2pq)</td>
<td>(m + \delta)</td>
<td>((q - p)a + (1 - 2pq)\delta)</td>
<td>(2(q - p)(1 - 2pq)a\delta + (1 - 2pq)^2\delta^2)</td>
</tr>
<tr>
<td>AA</td>
<td>(p^2)</td>
<td>(m + a)</td>
<td>(2q(a - p\delta))</td>
<td>(4q^2(a - p\delta)^2)</td>
</tr>
</tbody>
</table>

\[m - \alpha - m - (p - q)\alpha - 2pq\delta = \]
\[-p\alpha - (1 - q)\alpha - 2pq\delta = \]
\[-2p(\alpha - q\delta)\]  

(1.2)

Deviations for the other genotypes are listed in Table 1.2.

Next, we need to square the deviations from the mean. These are also listed in Table 1.2.

Finally, we must multiply the squared deviation from the mean by the frequency of the genotype and then sum over the three genotypes. Letting \(V_g\) denote the genetic variance, this gives the very messy equation

\[
V_g = 4p^2q^2(\alpha - q\delta)^2 + 2pq(q - p)^2\alpha^2 + 4pq(q - p)(1 - 2pq)\alpha\delta + 2pq(1 - 2pq)^2\delta^2 + 4p^2q^2(\alpha - p\delta)^2
\]

(1.3)

Without showing the tedious algebra, this reduces to

\[
V_g = 2pq[\alpha + (q - p)\delta]^2 + (2pq\delta)^2
\]

(1.4)

Equation 1.4 is composed of two terms. The first of these is called the additive genetic variance, usually denoted as \(V_a\),

\[
V_a = 2pq[\alpha + (q - p)\delta]^2
\]

(1.5)

If we plug the values for our example (\(p = 0.6\), \(\alpha = 0.5\), and \(\delta = 0.2\)) into this equation we arrive at \(V_a = .1016\).

The second term in 1.4 is called dominance genetic variance or \(V_d\)

\[
V_d = (2pq\delta)^2
\]

(1.6)

Our estimate of dominance variance for the example is \(V_d = .0092\).

The total genetic variable for our example is \(V_g = V_a + V_d = .1016 + .0092 = .1108\). Of this, \(.1016/.1108\) or 91.7% is additive and the rest (8.3%) is dominance.
1.2.1 Genetic Variance Components: Estimation (Graduate)

1.2.2 Heritability for a single locus (graduate)

In quantitative genetics, heritability is defined as the proportion of phenotypic variance predicted by (or attributable to) genotypic variance. Mathematically, it is a ratio with the genetic variance in the numerator and the phenotypic variance in the denominator, and it is custom to denote it as $h^2$. We have just seen that there are different types genetic variance for a single locus. Hence, there are different types of heritability.

The quantity $V_g$ in Equation 1.4 is the total genetic variance at the locus. Hence the quantity

$$h^2_b = \frac{V_g}{V_p}$$

is called the total heritability or, more often, broad sense heritability. Here, we denote is using the subscript $b$.

The ratio of the additive genetic variance to phenotypic variance is called narrow sense heritability or additive heritability. It is customary to represent this as just $h^2$ with no subscripts. One can also compute a dominance heritability by dividing $V_d$ by the phenotypic variance.

Why bother with these different types of heritability? The reason is that they become important for different types of prediction. For example, to predict the response to selection (either natural or artificial) or to predict the genetic values of offspring from parental values, we want to know about the additive genetic variance.

1.2.3 Meaning of the genetic variance components for a single locus (graduate)

Additive genetic variance measures the variance associated with an allelic substitution. That is, if I substituted allele $A$ for allele $a$ in the genotypes, then what is the variance associated with the change? This is not very intuitive, so let’s examine the mathematics behind this statement. Start with genotype $aa$. This genotype has no $A$ alleles so we will assign it a value of 0. With genotype $Aa$, we have substituted one $A$ for one $a$, so we give this genotype the value of 1. Finally, genotype $AA$ substitutes two $A$ alleles for $aa$ alleles, so its value is 2. We can estimate the additive statistics by performing a linear regression with these values as the predictor (or independent) variable and the phenotypic means for the genotypes (recall, these are also referred to as genetic values) as the predicted (or dependent) variable. Table 1.3 gives the notation used in this example where the relevant predictor variable is $X_{lin}$, the subscript standing for linear. For the moment, ignore the predictor variable $X_{quad}$.

The model for the additive regression is

$$\bar{Y} = \beta_0 + \beta_1 X_{lin} + \varepsilon$$

where $\beta_0$ is an intercept, $\beta_1$ is a slope and $\varepsilon$ is an error term.

At this point, a short digression is in order. There are two ways to perform this regression that differ according to the type of data set to be analyzed. In
Table 1.3: Schemata for a regression analysis at a single locus.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
<th>Phenotypic Mean</th>
<th>Predictors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>q²</td>
<td>$\bar{Y}_0$</td>
<td>$X_{lin}$ 0 0</td>
</tr>
<tr>
<td>Aa</td>
<td>2pq</td>
<td>$\bar{Y}_1$</td>
<td>1 1</td>
</tr>
<tr>
<td>AA</td>
<td>p²</td>
<td>$\bar{Y}_2$</td>
<td>2 0</td>
</tr>
</tbody>
</table>

the first case, we may have a data set of individual observations (e.g., individual people, rats, etc.). Here, the squared multiple correlations ($R^2$) for the regressions give will give us the genetic variance components. The second data set would consist of only three “observations”—the three genotypes and their frequencies, genetic values, and the codes given in Table 1.3. Here, the regressions must weight the means by the genotypic frequencies. The $R^2$ from these regressions gives the proportion of total genetic variance due to the genetic variance component.

Instead of algebraically deducing the regression, let us compute it using the computer language R. The relevant code for the regression is

```
additive <- lm(Ybar ~ Xlin, data=singleLocusExample, weights=frequency)
summary(additive)
```

Note that we must use the frequency variable to weight the regression because the genotypic frequencies are different. The relevant sections of output are

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1.1440   | 0.1920     | 5.958   | 0.106    |
| Xlin           | 0.4600   | 0.1386     | 3.320   | 0.186    |

Multiple R-squared: 0.9168

Hence, $\beta_0 = 1.144$ and $\beta_1 = 0.46$ so the predicted values of genotypes aa, Aa, and AA are, respectively, 1.144, 1.604, and 2.064. Notice that these are not the actual genotypic values of 1, 1.7, and 2.7. Why? Because there is dominance in the actual genetic values but we have only modeled the additive part. Note in the output that the multiple $R^2 = .917$. This is the same number we arrive at in Section 1.1.1 as the additive genetic variance divided by the total genetic variance.

What does this tell us? The regression demonstrates that the additive genetic variance is the variance associated with placing the best-fitting straight line through the three genetic values. That line is illustrated by the solid straight line in Figure 1.3.

What about dominance variance? There are several ways to estimate this but the simplest way is to create a new predictor variable that has a value of 0 for
the homozygotes but 1 for the heterozygote. This is variable $X_{quad}$ (subscript quad for quadratic) in Table 1.3. We then perform a multiple regression with the genetic values as the dependent variable and $X_{lin}$ and $X_{quad}$ as the two predictor variables. The regression model is now\(^1\)

$$\bar{Y} = \beta_0 + \beta_1 X_{lin} + \beta_2 X_{quad} \quad (1.9)$$

The relevant R code and output are

```r
total <- lm(Ybar ~ Xlin + Xquad , data=singleLocusExample , weights=frequency)
summary(total)
```

Coefficients:

|         | Estimate | Std. Error | t value | Pr(>|t|) |
|---------|----------|------------|---------|----------|
| (Intercept) | 1.0      | NA         | NA      | NA       |
| Xlin     | 0.5      | NA         | NA      | NA       |
| Xquad    | 0.2      | NA         | NA      | NA       |

Multiple R-squared: 1

Notice that the coefficient for the linear term, i.e., $\beta_1$, now equals the value of $\alpha$ from Section 1.1.1–0.5. Similarly, the coefficient for the dominance term, $\beta_2$, equals $\delta$ from the example–0.2. Notice also that the $R^2$ equals 1.0 because we are predicting three observed quantities (the three genetic values) from three

\(^1\)Because we are predicting three data points from three unknowns (i.e., the $\beta$s in the equation), there will be a perfect fit to the data. Hence the equation does not contain an error term.
unknowns (the three $\beta$s). Hence, the predicted values for the three genotypes will equal the observed values and a plot of the predicted values will be part of a quadratic equation (i.e., a parabola) as illustrated by the dashed line in Figure 1.3. The proportion of total genetic variance due to dominance equals this $R^2$ minus the $R^2$ from the additive model or $1 - .917 = .083$.

This procedure illustrates a very important property of genetic variance components—namely, they are estimated hierarchically. We estimate additive genetic variance first. We then statistically remove that from the total genetic variance and let the remainder be defined as dominance variance. As a result, once can have dominant genes that have very little dominance variance (see text box).
Gene Action and Genetic Variance Components

Gene action is required for a variance component. For example, without dominant gene action, there can be no dominance variance. However, genetic variance components are a function of both gene action and genotypic frequencies. Consequently, one can have strong gene action, but if the genotypic frequencies are just right, then the variance component associated with it can be very small. To illustrate this consider the two equations given in the text to compute the additive and dominance variance at a single locus from the regression parameters:

\[ V_a = 2pq [\alpha + (q - p)\delta]^2 \]

and

\[ V_d = (2pq\delta)^2 \]

Let us assume that allele \( A \) shows complete dominance to allele \( a \) so that \( \delta = \alpha \). Then the ratio of \( V_a \) to \( V_d \) is

\[ \frac{V_a}{V_d} = \frac{2pq [\alpha + (q - p)\alpha]^2}{(2pq\alpha)^2} \]

which reduces to

\[ \frac{V_a}{V_d} = 2 \frac{q}{p} \]

Even though we have modeled a completely dominant gene, this equation tells us that the ratio of additive to dominance variance depends only on the allele frequencies! When allele frequencies are even (i.e., \( p = q \)) then the ratio is 2 and we will have twice as much additive variance as dominance variance. As the frequency of the recessive allele \( a \) increases, \( q \) becomes larger and larger and there is more and more additive variance relative to dominance variance. This tells us that rare dominant alleles have large additive variance relative to their dominance variance.

As \( q \) becomes smaller and smaller relative to \( p \), the ratio will get less than 1 and approach 0. Thus, loci with rare recessive alleles have large dominance variance relative to their additive variance.

1.2.4 More than two alleles (graduate)

Most genes have many more than two alleles. How does one model this situation? The answer is surprisingly similar to the two allele case. We construct predictor variables for the additive effects of allele substitutions and then other predictors that indicate heterozygotes. They key is that if there are \( k \) alleles, then there will be \((k - 1)\) predictor variables for the additive variables and
Table 1.4: Coding for a locus with three alleles.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>(X_{lin1})</th>
<th>(X_{lin2})</th>
<th>(X_{quad1})</th>
<th>(X_{quad2})</th>
<th>(X_{quad3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1A_1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(A_1A_2)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(A_1A_3)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(A_2A_2)</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(A_2A_3)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(A_3A_3)</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\(0.5k(k-1)\) predictors for the dominance variables. We actually followed this rule in the two allele case from Table 1.3. Here \(k = 2\), so there was one additive predictor and one dominance predictor.

As an example, consider a locus with three alleles, \(A_1\), \(A_2\), and \(A_3\). Because \(k = 3\), we will have two additive variables. Let us ignore the first allele and construct the first additive or linear variable as number of \(A_2\) alleles in a genotype and the second as the number of \(A_3\) alleles in the genotype. Table 1.4 gives this coding and the respective names of the variables as \(X_{lin1}\) and \(X_{lin2}\).

For the dominance, we construct a variable for each possible heterozygote in the sample. With three alleles, there are three possible heterozygotes–\(A_1A_2\), \(A_1A_3\), and \(A_2A_3\). Hence we create three variables. For each variable, if the genotype is the relevant heterozygote then the value of the variable is 1; otherwise, it is 0. Table 1.3 shows the three dominance variables for this example.

The proportion of total genetic variance that is additive will equal the squared multiple correlation \(R^2\) from the regression of the genotypic values on the two additive variables or

\[ \bar{Y} = \beta_0 + \beta_1 X_{lin1} + \beta_2 X_{lin2} + \varepsilon \]  

(1.10)

The proportion that is dominance variable will equal 1 minus this \(R^2\).

1.2.5 Two locus model (graduate)

In a model of two loci, we will have an additive effect for the first locus, an additive effect for the second locus, a dominance effect for the first locus and a dominance effect for the second locus. In keeping with the hierarchical estimation of variance components, we would first fit a regression with the only the additive effects. Call this Model 1. Its equation is

\[ \text{Model 1}: \bar{Y} = \beta_0 + \beta_1 X_{Add1} + \beta_2 X_{Add2} + \varepsilon \]  

(1.11)

The \(R^2\) from this would be the additive genetic variance (if the data consisted of individuals) or the proportion of total genetic variance that is additive (if the data were phenotypic means).
1.2. GENETIC VARIANCE COMPONENTS

Figure 1.4: An example of a non interactive model (left panel) and genetic epistasis (right panel).

Next we would fit a model with the two additive terms and the two dominance terms or

Model 2: \[ \bar{Y} = \beta_0 + \beta_1 X_{Add1} + \beta_2 X_{Add2} + \beta_3 X_{Dom1} + \beta_4 X_{Dom2} + \varepsilon \]  

We would then subtract the $R^2$ from the first, additive model from the $R^2$ from Model 2. The result would be the dominance variance (if the data were from individual observations) or the proportion of genetic variance due to dominance variance (if the data were means). In keeping with the

In addition to additive and dominance effects, a two locus model may also contain another source of genetic variance—that due to the interaction between genes or epistasis. Before exploring this issue, however, we must be clear about definitions. The term “interaction” has two meanings. The first is a fuzzy definition that implies that both loci contribute to the phenotype. The second is a precise, statistical definition—namely, the effect of one locus on the phenotype depends on the genotype at the second locus. Epistasis applies only to the statistical definition.

Figure 1.4 provides an example. The panel on the left illustrates a case of no gene by gene interaction. Here, the effect of the A locus does not depend on the value of the B genotype. Relative to $A_1A_1$, genotype $A_1A_2$ always increases the genetic value by 2.5 units across all three genotypes at the B locus. Similarly, genotype $A_2A_2$ increases the response by 3 units across all genotypes at the B locus.

Geometrically, the hallmark of a non interactive model is that the lines for the B locus all have the same shape even though they may differ in elevation.

The right-hand panel illustrates epistasis. Here the effect of A genotypes depend strongly on the B locus. With genotype $B_1B_1$, the A locus has no effect on the phenotype. Relative to $A_1A_1$, genotype $A_1A_2$ increases the genetic value by 1 unit when the B genotype is $B_1B_2$ but by 2.5 units when it is $B_2B_2$. Geometrically, the lines for the B locus do not have the same shape.

To model epistasis, we must recognize that there are different forms of interaction. With two loci, there may be an interaction between the additive effects
CHAPTER 1. QUANTITATIVE GENETIC VARIANCE COMPONENTS: ESTIMATION (GRADUATE)

Table 1.5: Coding for the analysis of epistasis for two loci.

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
<th>Add1</th>
<th>Add2</th>
<th>Dom1</th>
<th>Dom2</th>
<th>AA</th>
<th>A1D2</th>
<th>A2D1</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁A₁</td>
<td>B₁B₁</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₁A₁</td>
<td>B₁B₂</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₁A₂</td>
<td>B₁B₂</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₁A₂</td>
<td>B₁B₁</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>A₁A₂</td>
<td>B₁B₂</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A₁A₂</td>
<td>B₂B₂</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A₂A₂</td>
<td>B₁B₁</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₂A₂</td>
<td>B₁B₂</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A₂A₂</td>
<td>B₂B₂</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

at locus 1 and the additive effects at locus 2. This is called additive by additive epistasis. In a regression model the coefficients for the genotypes may be found by multiplying the coefficients for the additive effects at locus one by the coefficients for the additive effects at locus two.

To illustrate, suppose that we used contrast codes for the additive effects. Here, the heterozygote is assigned a numerical value of 0; one of the homozygotes, a value of -1; and the other homozygote, a value of 1. Table 1.5 illustrates these codes. The column labeled “AA” gives the codes for the additive by additive epistasis. You should verify that this code is formed by multiplying the additive codes for locus A by those for locus B.

We now fit Model 3 to the data using the equation

\[
\bar{Y} = \beta_0 + \beta_1 X_{\text{Add1}} + \beta_2 X_{\text{Add2}} + \beta_3 X_{\text{Dom1}} + \beta_4 X_{\text{Dom2}} + \beta_4 X_{\text{AA}} + \varepsilon
\]  

(1.13)

By subtracting the \( R^2 \) for this model from the \( R^2 \) from Model 2, we arrive at the additive by additive epistatic variance (or the proportion of total genetic variance that is additive by additive epistatic variance).

The next type of epistasis occurs between the additive effects at one locus and the dominance effects at the other locus—additive by dominance epistasis. Here, we have two terms. The first will be the additive effects at the A locus and the dominance effects at the B locus. The column labeled “A1D2” in Table 1.5 gives the code for this effect. It is formed by multiplying the additive code for the A locus (“Add1”) by the dominance code for the B locus (“Dom2”).

The second code is for the dominance effect at the A locus and the additive effects at the B locus. This is given by column “A2D1” in Table 1.5 and it is formed as the product of “Add2” and “Dom1.”

Out next model—Model 4—adds these two terms to the regression equation

\[
\bar{Y} = \beta_0 + \beta_1 X_{\text{Add1}} + \beta_2 X_{\text{Add2}} + \beta_3 X_{\text{Dom1}} + \beta_4 X_{\text{Dom2}} \beta_4 X_{\text{AA}} + \beta_3 X_{\text{AA}} + \beta_2 X_{\text{A1D2}} + \beta_1 X_{\text{A2D1}} + \varepsilon
\]  

(1.14)

The additive by dominance epistatic variance (or the proportion of total variance
Table 1.6: Model $R^2$'s and estimates of variance components.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>Component</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>.410</td>
<td>Additive</td>
<td>.410</td>
</tr>
<tr>
<td>Model 2</td>
<td>.426</td>
<td>Dominance</td>
<td>.016</td>
</tr>
<tr>
<td>Model 3</td>
<td>.437</td>
<td>Add x Add</td>
<td>.011</td>
</tr>
<tr>
<td>Model 4</td>
<td>.446</td>
<td>Add x Dom</td>
<td>.009</td>
</tr>
<tr>
<td>Model 5</td>
<td>.450</td>
<td>Dom x Dom</td>
<td>.004</td>
</tr>
<tr>
<td>Total</td>
<td>.450</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

due to additive by dominance epistasis if means are analyzed) is derived by subtracting the $R^2$ for Model 3 from the $R^2$ from this model.

The final epistatic term is dominance by dominance epistasis. Here, we arrive at the code by multiplying the dominance code for the $A$ locus by the dominance code for the $B$ locus giving column “DD” in Table 1.5. We now add this term to the model giving

$$
\bar{Y} = \beta_0 + \beta_1 X_{Add1} + \beta_2 X_{Add2} + \beta_3 X_{Dom1} + \beta_4 X_{Dom2} + \beta_4 X_{AA}
+ \beta_4 X_{AA} + \beta_5 X_{A1D2} + \beta_6 X_{A2D1} + \beta_7 X_{DD} + \varepsilon
$$

Once again, the $R^2$ for this model less the $R^2$ from Model 4 gives the dominance by dominance variance component.

1.2.5.1 A numerical example (graduate)

A numerical example will help to illustrate the procedure. Let us take the data that generated the epistatic panel in Figure 1.4 and calculate the variance components. Here, it was assumed that the overall effect of the two loci on the phenotype accounted for 45% of the phenotypic variance and that the frequencies of alleles $A_1$ and $B_1$ were respectively 0.6 and 0.3.

Table 1.6 present the results of the regressions. If you glance down the column giving the estimates of the variance components, you will find it striking how the magnitude of the components decreases as one goes from the additive variance to the dominance by dominance epistatic variance. This is not unique to this example. Instead, it derives from the fact that genetic variance components are estimated hierarchically.
It is important to reflect on the hierarchical nature of the estimated genetic variance components. The first component is the additive genetic variance and regression procedure will maximize the amount of this variance in explaining the dependent variable. After dominance is extracted, regression will perform another maximizing step to arrive at dominance variance. Then, having already accounted for a large chunk of the genetic variance, it will try to maximize the additive by additive epistatic variance.

As a result of this process, the largest components will be extracted first and the components will have a tendency to get smaller and smaller with each regression. The result is similar to the text box on “Gene Action and Genetic Variance Components” for a single locus–namely, one may observe epistatic gene action but not translate into noticeable epistatic variance.

In summary, it is important to recognize that biological gene action and statistical variance components are not completely interlocked. Even though there may be considerable gene-gene interaction at the biological level, there may be very little gene-gene interaction variance at the statistical level.

In practice, there are few instances in which one wants to calculate the various components of epistatic variance. Indeed, the importance of this exercise was to impress on the reader the hierarchical nature of estimating genetic variance components. It is crucial to recognize the difference between biological gene action and the variance components associated with that gene action.

1.2.6 More than two loci (graduate)

Technically, calculation of variance components for more than two loci follows the logic outlined above—it is just that the number of predictor variables gets very large. For example, with four bi-allelic loci, there will be four additive genetic variables, one for each locus. There will be six dominance variables, one for each of the six possible heterozygotes. Codes for epistatic variables must include all two-way interactions, all three-way interactions, and all four-way interactions. Each interaction will also have separate variables for the various combinations of additive and dominance components. For example, there will be six variables used to predict additive by additive epistasis–additive code for locus 1 times additive code for locus 2, additive code for locus 1 times additive code for locus 3, etc. There will be 12 variable for additive by dominance epistasis and 15 for dominance by dominance.

Current genetic research suggests that there may be hundreds, perhaps thousands, of DNA sections that contribute to individual differences in behavior. Even if only a handful of these are the subject of analysis, it would be prohibitive to create all of the codes necessary to partition epistatic variance. Instead, one could compute the means for all the genotypes in the sample. The pooled variance within the genotypes would equal the variance not due to this locus; this
1.3 SOME IMPORTANT POPULATION GENETIC CONCEPTS

We cannot survey all relevant aspects of population genetics. Instead, let us overview a few concepts relevant to human genetics and behavior. For many years, the “bible” has been Crow & Kimura (X.X). Other relevant texts are

1.3.1 Hardy-Weinberg equilibrium

In diploid populations when: (1) the population is large with no significant migration or immigration; (2) there is no selection influencing the gene in question;
CHAPTER 1. QUANTITATIVE POPULATION GENETIC CONCEPTS

(3) mating is random with respect to the gene under study, then two phenomena will occur. The first states that the frequency of a genotype in the population will equal the expected frequency under the condition that the alleles in that genotype pair at random, i.e., they follow the laws of probability applied to independent random events. As an example, consider a bi-allelic loci with alleles $A$ and $a$ with respective frequencies of $p$ and $q = (1 - p)$. When the assumptions hold, then the frequency of genotype $AA$ is the probability of randomly picking an $A$ allele from a “hat” with a very large number of alleles and then reaching into the hat a second time and also picking an $A$. The probability of picking an $A$ allele is $p$ and, given that the second pick is independent, the probability of picking a second $A$ is also $p$. Hence, the probability of genotype $AA$ is the product of these probabilities or $p^2$. By similar logic, the expected frequency of genotype $aa$ is $q^2$.

The expected frequency of the heterozygote also follows the rules governing the probability of independent events, but differs because there are two different ways of getting an $Aa$ genotype. First, we could reach into the hat, pick an $A$ and then reach in again and pick an $a$. The probability of this event is $pq$. Secondly, we could first pick an $a$ and then pick an $A$. Thus probability is $qp$. The total probability is the sum of these two probabilities, so the expected frequency of $Aa$ equals $pq + qp = 2pq$. In sum, the frequency of genotypes $aa$, $Aa$, and $AA$ will equal the terms in the expansion of the binomial $(q + p)^2$ or $q^2$, $2pq$ and $p^2$.

The second phenomena is that the genotypic frequencies will remain the same from one generation to the next. There are complicated ways of proving this mathematically, but it is easiest to see by treating alleles picked out of a big hat as alleles in gametes generated from random parents. The probability of two random gametes each containing an $A$ equals $p^2$, etc.

When this state of affairs occurs, then the gene is said to be in Hardy-Weinberg equilibrium often abbreviated as $H-W$ equilibrium. Note that the equilibrium condition applies to individual sections of DNA and not to the whole genome. Some loci may be in $H-W$ equilibrium while other loci are not in equilibrium.

The extension to multiple alleles is trivial if one follows the principles outlined above. Let the alleles be denoted as $A_1, A_2, \ldots, A_k$ with respective probabilities of $p_1, p_2, \ldots, p_k$. The the probability of genotype $A_iA_j$ is $p_i^2$ when $i = j$ and $2p_ip_j$ otherwise.

The usual test for $H-W$ equilibrium is a $\chi^2$ goodness-of-fit test. The layout is given in Table 1.7. We begin by obtaining an estimate of $p$ from the observed data. With a sample size of 986 there are $2 \times 986 = 1,972$ alleles. There are 372 $AA$ individuals so these contribute $2 \times 372 = 744$ $A$ alleles. There are 402 heterozygotes each contributing one $A$ allele. Hence the frequency of $A$ in the population is $p = (744 + 402)/1972 = 0.581$.

Next we compute the expected number under the hypothesis of $H-W$ equilibrium. For genotype $AA$, this is $p^2 \times 986 = 332.8$; for $Aa$, it is $2pq \times 986 = 480.1$; and for $aa$, it is $q^2 \times 986 = 173.1$. 17
1.3. SOME IMPORTANT POPULATION GENETIC CONCEPTS

CHAPTER 1. QUANTITATIVE

Table 1.7: Goodness-of-fit test for H-W equilibrium.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Type:</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>aa</td>
<td>Aa</td>
</tr>
<tr>
<td>Observed</td>
<td>212</td>
<td>402</td>
</tr>
<tr>
<td>Expected</td>
<td>173.1</td>
<td>480.1</td>
</tr>
</tbody>
</table>

Figure 1.5: Notation for selection for a single locus.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
<th>Mean</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>$q^2$</td>
<td>$m - \alpha$</td>
<td>$s_{aa}$</td>
</tr>
<tr>
<td>Aa</td>
<td>$2pq$</td>
<td>$m + \delta$</td>
<td>$s_{Aa}$</td>
</tr>
<tr>
<td>AA</td>
<td>$p^2$</td>
<td>$m + \alpha$</td>
<td>$s_{AA}$</td>
</tr>
</tbody>
</table>

Next we compute the $\chi^2$ statistic using the generic formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where $O_i$ is the observed number and $E_i$, the expected number for the ith cell. For the present example, $k = 3$, so the formula yields

$$\chi^2 = \frac{(212 - 173.1)^2}{173.1} + \frac{(402 - 480)^2}{480.1} + \frac{(372 - 332.8)^2}{332.8} = 26.06$$

This statistic is highly significant, so the population is not in H-W equilibrium for this locus.

What the test does not tell us is why the population departs from equilibrium. Comparison of the observed with the expected frequencies in Table 1.7 suggests that there are more homozygotes than expected. One possible reason for this is that the population is actually a mixture of two subpopulations, one with a high frequency of $A$ alleles and the other with a high frequency of $a$ alleles.

1.3.2 Selection at a single locus (graduate)

What happens if there is natural selection at a locus? Here we use the model developed for a single locus above in Section 1.5. Let us begin with a population in H-W equilibrium and examine the effects of selection in the offspring generation. We must now add selection coefficients or algebraic quantities that stand for likelihood of a genotype contributing to the next generation. There are several ways of doing this. Here, we just assign an algebraic quantity $s$ to each genotype. The setup is shown in Table.

The $s$s in Table 1.5 operate as “weights” that are applied to the gametes created by each genotype. Hence, the frequency of gametes containing allele $A$
CHAPTER 3. QUANTITATIVE POPULATION GENETIC CONCEPTS

will be

\[ 2p^2s_{AA} + 2pq s_{Aa} \]

This quantity, however, must normed by the frequency of all gametes so the actual frequency of allele A among all transmitted gametes will be

\[ \frac{2p^2s_{AA} + 2pq s_{Aa}}{2(p^2s_{AA} + 2pq s_{Aa} + q^2s_{aa})} = \frac{p(ps_{AA} + q s_{Aa})}{p^2s_{AA} + 2pq s_{Aa} + q^2s_{aa}} \]  \hspace{1cm} (1.19)

and the frequency of allele a in the offspring will be

\[ \frac{q(qs_{aa} + p s_{Aa})}{p^2s_{AA} + 2pq s_{Aa} + q^2s_{aa}} \] \hspace{1cm} (1.20)

To view how selection works, we need to implement special cases by giving numerical values to the selection coefficients. One of the easiest cases is a lethal autosomal dominant. Let A denote the dominant allele. Here, \( s_{AA} = s_{Aa} = 0 \) and the denominator in Equation 1.19 becomes 0. Hence, the lethal allele is eliminated in the next generation. It is for this reason that lethal dominant disorders are never transmitted unless, like Huntington’s disease or the Mendelian forms of Alzheimer’s disease, they have delayed onset so that reproduction occurs before the onset of the disorder. Any dominant lethal disorder will be the result of a new mutation. An example of such a lethal is Hutchinson-Gilford progeria which causes premature aging.

Let us now examine the case of selection against a recessive. and let \( s_{AA} = a_{Aa} = 1 \). Then \( s_{aa} \) will be some number less than 1. The frequency of allele a in the next generation will be

\[ \frac{q(qs_{aa} + ps_{Aa})}{p^2 + 2pq + q^2s_{aa}} \]

If the condition is lethal, then \( s_{aa} = 0 \) and the numerator reduces to \( pq \). Note that this quantity will never equal 0. Instead, \( q \) will get smaller and smaller over subsequent generations becoming very close to, but never equal to, 0.\(^2\)

Imagine a recessive allele with a high frequency that suddenly becomes lethal. Such an abrupt transition is unlikely to occur in nature but it will serve to illustrate some principles about selection. Figure 1.6 plots the frequency of the recessive allele over 25 generations assuming that its initial frequency is 0.9. (For the moment, ignore the lines for the additive and dominance variance.) There is a very rapid response to selection in the first few generations. After that, the pace of change gradually slows until by, say, generation 20, the rate of change is very minimal.

\(^2\)Technically, the allele frequency will decline until it reaches a value where the loss in affected homozygotes is replaced by new mutations. This is called a mutation-selection balance and under simple models occurs when the frequency of the disorder is approximately \( \sqrt{\mu/(1-s)} \) where \( \mu \) is the mutation rate for changing a normal allele into a deleterious recessive allele.
1.3. SOME IMPORTANT POPULATION GENETIC CONCEPTS

CHAPTER 1. QUANTITATIVE

There is a logical reason for this pattern. As the recessive allele becomes rare, the frequency of the recessive condition becomes vanishing rare relative to the frequency of the unaffected heterozygote. For example, when $q = .01$, then the prevalence of the condition is 1 in 10,000 births. Yet, the frequency of the heterozygote rounds off to 2% of the general population. That is why it is so hard to eliminate a recessive condition. The overwhelming majority of organisms carrying the allele are unaffected heterozygotes.

1.3.3 Selection and continuous traits (graduate)

To calculate the effects of selection on a continuous trait, it is necessary to define three mathematical functions: (1) the distribution of the trait at a point in time (or generation); (2) a selection function; and (3) a transmission model. Assuming a polygenic background and a normal distribution of environmental effects, then it is reasonable to assume that the phenotype will be normally distributed. That takes care of the first function.

A selection function is a mathematical function of phenotypic values and parameters. Give the parameters, the selection function gives the probability that a person with a specific phenotypic value will contribute a gamete (or offspring) to the next generation. If $f(X)$ is the equation for the normal curve where $X$ is the phenotype and $s(X)$ is the selection function for the phenotype, then the proportion of the population that reproduces equals the integral

$$p_r = \int_{-\infty}^{\infty} f(X)s(X)dX$$

(1.21)

and the distribution of the phenotype among those reproducing is

$$f_r(X) = f(X)s(X)/p$$

(1.22)

The problem with selection functions is that it may not be possible to arrive at a closed-form solution—or even a close approximation for that matter—to the quantity $f(X)s(X)$. The situation can be further complicated when the phenotype $X$ is not the direct object of selection but instead a trait that is correlated with a phenotype that is the direct object of selection.

The final function is transmission which
1.3.3.1 Fisher’s fundamental theorem of natural selection

In 1930, R.A. Fisher published a treatise on

1.4 References


Plutynski, A. (2006), What was Fisher’s fundamental theorem of natural selection and what was it for? Studies in History and Philosophy of Biological and Biomedical Sciences, 37:59-82