VARIATION AND VARIANCE

The essence of Quantitative Genetics is the exploration of biological variation, the main goal being to create a framework for understanding and quantifying how differences genes and environment influence this variation. In order to accomplish this task, it is first necessary to define what is meant by "biological variation." It is certainly easy enough to look at the world and see remarkable examples of it, but how do we define it? For instance, which of the two populations of beetles shown in Figure 1.1 displays the greatest biological variation? The most commonly selected population would likely be A, containing beetles displaying anywhere from no spots to six spots per wing. This certainly appears to be the population showing the greatest diversity for this trait. However, is *diversity* the same as *variation*? To answer this, it is necessary to define variation, or more accurately, to define how we *measure* it. In Quantitative Genetics, we measure variation via **variance**.

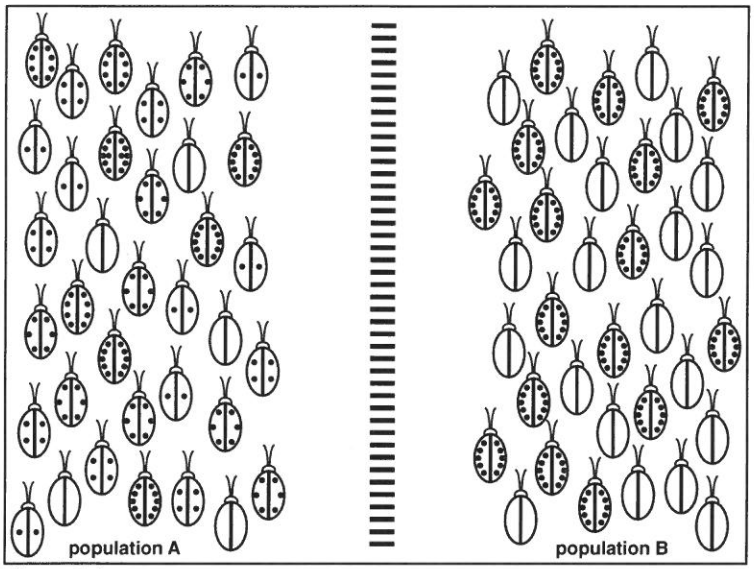


Figure 1.1

To understand quantitative genetics, you must understand variance

Variance is a measure of how far observations are expected to be from the population mean. It is a mathematical concept and is formally defined with an equation. This should in no way cause (or excuse) a lack of intuitive understanding of the concept, however. To truly understand variance, it is necessary to introduce the concept of a random variable. It is worth mentioning that a useful, though somewhat counter-intuitive feature of Quantitative Genetics is that, at its core, it is actually discrete. While the traits we measure are quantitative in nature, the genetics, or more specifically the genotypes we are interested in ascribing those traits to, are discrete entities. One implication of this is that to understand the quantitative genetics model, we can focus on discrete random variables.

Discrete Random Variables

Consider that you perform an experiment in which all possible outcomes can be enumerated. For example, consider an experiment in which an individual is randomly selected from a population, and its genotype at a given locus (the outcome of the experiment) is recorded. If the locus has two alleles segregating in the population, for example "B" and "b", you can enumerate all possible genotypes (i.e. all possible outcomes) as "BB", "Bb", and "bb". Another example of an experiment for which all possible outcomes can be enumerated is Mendel's experiment of crossing pea plants with different color seeds

(yellow and green) and observing the colors of seeds among the offspring. The possible outcomes of this experiment would be "yellow" and "green". A third example of an experiment for which all possible outcomes can be enumerated is observing the number of spots on the wing of a beetle selected at random from one of the populations shown in Figure 1.1.

For our purposes, we can describe a **discrete random variable** as a statistical tool for modeling this type of experiment. The values that the random variable can take on are numeric, and map to the possible outcomes of the experiment. A simple example of such a random variable is one that models the outcome of the experiment in which a beetle is selected at random from one of the populations in Figure 1.1 and the number of spots on its wing recorded. The random variable, which we can denote as X, in this case can take on the value representing the number of spots observed. In the case of experiments with non-numeric outcomes, such as observing the seed color of an offspring from one of Mendel's pea plant crosses, outcomes must be mapped to numeric values. For example, we could define a random variable X that takes on the value 0 if the outcome of the experiment is "green" and 1 if the outcome is "yellow".

Also useful in modeling this type of experiment is the **probability distribution**, or the set of probabilities that are associated with the values that the random variable can take on. In the case of the beetles, the probability of randomly selecting an individual with a particular number of spots is equal to the proportion of beetles with this number of spots in the overall population (if 20% of beetles in the population have five spots per wing, then the probability of the random variable taking on the value 5 is 0.20). For a more complicated example, consider a population in which most beetles do not have spots. A few beetles in this population, however, are homozygous for a somewhat rare mutation that causes them to display four spots. If the frequency of the rare allele in the population is 10%, and Hardy-Weinberg genotype proportions exist in the population, then the probability of randomly sampling a beetle with four spots is 0.01 (the allele frequency squared). If X is a random variable that models the number of spots of a randomly sampled beetle, the possible values X can take on are 0 and 4, with probabilities 0.99 and 0.01, respectively.

Here, the value a random variable X can take on will be denoted by x, and the probability that $X = x$ denoted by p (this can be written as $\Pr[X=x] = p$). In the example just described, the possible values X can take on are x=0 or x=4, and $\Pr[X=0] = 0.99$ and $\Pr[X=4] = 0.01$.

Expected Values

It is crucial to understand the concept of an expected value in order to understand variance, and thus in order to have a good grasp of the basics of Quantitative Genetics. The simplest definition of an expected value is a "long-term" mean: if you repeated an experiment modeled by a random variable X an infinite number of times, and observed each of these infinite number of outcomes, the mean of these observations would be that random variable's expected value. If you can only repeat the experiment a finite number of times and thus derive a finite number of observations, then you can only calculate the *sample* mean, which is an *estimate* of the expected value. It is worth noting that the expected value of a random variable may be a value that cannot be observed in a single observation. For example, if random variable X represents the toss a fair six-sided die, the values it can take are $x = 1, 2, \ldots, 6$ (all with equal probability of 1/6). The expected value of this random variable is 3.5, which is not a possible die toss.

We write expected values using the notation $E[\ ]$, so that the expected value of the random variable X is $E[X]$. To calculate the expected value for a discrete random variable, X, you need to know two things. One is the values X can take on, and the other is the probabilities of seeing those values. The calculation is quite simple: $E[X] = \sum_i x_i p_i$, where the x_i are all the values X can take on, and the p_i are the probabilities of seeing these values. For instance, in the example of the six-sided fair die, each side has a one in six chance of appearing, so that

$$E[X] = \sum_{i=1}^{6} \frac{1}{6} i = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

While dice provide a useful discrete probability tool, they are not very meaningful in a genetics context. As a simple, more relevant example, consider a genetic locus with two alleles, B and b. Individuals with a BB or Bb genotype are tall, say 14 feet, and individuals with a bb genotype are short, say 11 feet. Consider the experiment of crossing two heterozygous parents and observing the height of the offspring of this cross. If X is the random variable that represents the height of an offspring, what is $E[X]$?

Solution:
Remember that to calculate an expected value we need to know the values the random variable can take on and the probabilities of seeing those values. Here, the possible values are 11 feet and 14 feet. There are two types of 14 foot tall organisms: those with a BB genotype and those with a Bb genotype. There is one type of 11 foot tall individuals, those with a bb genotype. The probability of seeing a BB genotype from this cross is 1/4, the same as the probability of seeing a bb genotype. The probability of a Bb genotype is 1/2. So, $E[X] = \sum_i x_i p_i = 14 \cdot \frac{1}{4} + 14 \cdot \frac{1}{2} + 11 \cdot \frac{1}{4} = 13.25$ feet.

How does an expected value compare to a sample average? Using the same example as above, consider a "perfect" sample of eight offspring. Here "perfect" means that the sample exactly reproduces the model probabilities: of eight offspring, two are BB and 14 feet tall, four are Bb and 14 feet tall, and 2 are bb and 11 feet tall (achieving the classic ratio expected from a cross of two heterozygotes, 1/4:1/2:1/4). For this sample, the sample mean is:

$$\frac{(14+14)+(14+14+14+14)+(11+11)}{8}$$

$$= \frac{14 \cdot 2 + 14 \cdot 4 + 11 \cdot 2}{8}$$

$$= 14 \cdot \frac{1}{4} + 14 \cdot \frac{1}{2} + 11 \cdot \frac{1}{4}$$

which is the same as the calculation for the expected value, above. In this way, the expected value can be described as being the mean of a "perfect" sample. Of course only certain samples sizes are capable of yielding a "perfect" sample; if we had a different sample size, for instance 10 offspring, then a "perfect" sample would not be possible, and the sample mean could not equal the expected value.

In summary, even though the equation for an expected value does not on the surface look like the equation for a sample mean, the expected value is, indeed, a mean. This value is often represented by the parameter μ:

$$\mu = E[X].$$

Variance

As mentioned previously, variance (often represented as σ^2) is a measure of how far away, on average, random observations are expected to be from their population mean. It is formally defined through an equation:

$$Var[X] = \sigma^2 = E[(X - E[X])^2]$$

for the random variable X. This equation contains two expected values. The inner one, $E[X]$ is simple: it is exactly as described above, and we can substitute in the parameter μ for ease of notation:

$$\sigma^2 = E[(X-\mu)^2].$$

The remaining expected value expression may look somewhat complicated, but can be understood intuitively. First consider the concept of the expected value of a variable squared: $E[X^2]$, which is simply $\sum_i x_i^2 p_i$ (p_i is the probability X can take on value x_i, but instead of multiplying the value of x_i by p_i, we multiply the value of x_i^2 by p_i). For example, for a fair six-sided die we know that X can take on the values 1, 2, …, 6, all with probability 1/6. This means that X^2 can take on the six values 1, 4, 9, 16, 25, and 36, each also with probability 1/6. So, from the equation of an expected value,

$$E[X^2] = \sum_i^6 i^2\left(\frac{1}{6}\right) = 1\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 9\cdot\frac{1}{6} + 16\cdot\frac{1}{6} + 25\cdot\frac{1}{6} + 36\cdot\frac{1}{6} = 15\frac{1}{6}.$$

$E[X^2]$ is simply the mean value of X^2 in an infinite sample (or in a "perfect" sample).

Extending from this, $\sigma^2 = E[(X-\mu)^2]$ is the mean of $(X-\mu)^2$ in an infinite sample. The inner part, $X-\mu$, represents the distance of an observed value of X from the mean; $(X-\mu)^2$ is just that value squared. So, $\sigma^2 = E[(X-\mu)^2]$ is simply how far away you can expect, on average, X to be from its mean, squared. Intuitively, if on average observations are expected to be far from their mean, variance is high. If you expect values to be, on average, close to their mean, variance is small.

The Effect of Probabilities on Variances

We have mentioned that an expected value of a random variable depends on two things, the values that the random variable can take on, and the probabilities that these values will occur. Since variance is an expected value, it is clear that the same thing is true for it. The effect of probabilities on expected values, including on variance, are important but often overlooked. Consider what happens if we exchange our fair die for a loaded die; for example, one which lands on the number six 50% of the time, and lands on the other numbers 10% of the time for each of them. If X is the random variable that represents a toss of this loaded die, its expected value will no longer be 3.5; instead, it will larger (since the probability of seeing the largest value, 6, is high). Variance will also be affected by this change, as now several things have changed: the mean, the distance between an observed toss and the mean, and the probability of seeing each of these distances. In a biological context the relationship between variances and probabilities is important, as often probabilities change from population to population. For instance, in one population the probability of encountering a BB genotype (the frequency of that genotype) is likely to be different than it is in a another population, even if the height of individuals with those genotypes does not change across populations. In our previous example of the effect of a locus on height of an individual, the mean height and the variance in height both depend on the genotype frequencies.

Exercise 1.1: Calculate the mean height and variance in height for the following populations. For all populations, the height of individuals with BB and Bb genotypes is 14 feet, and the height of individuals with a bb genotype is 11 feet.

	genotype frequency		
pop.	BB	Bb	bb
1	0.01	0.18	0.81
2	0.25	0.50	0.25
3	0.81	0.18	0.01

Solutions:

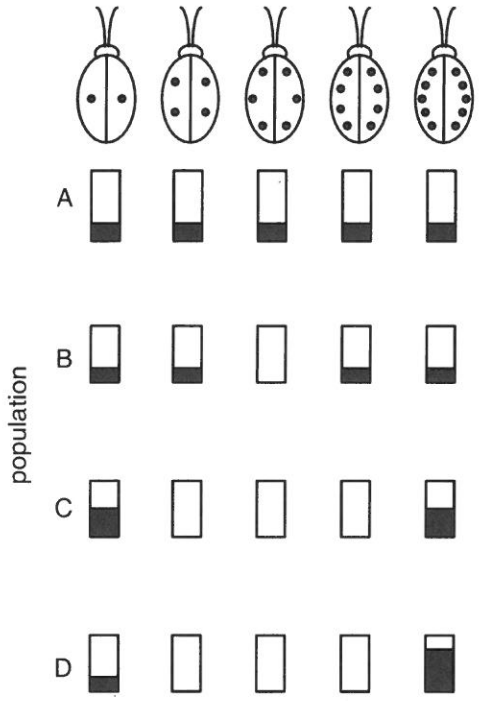
pop.	μ	σ^2
1	11.57 feet	1.39 feet^2
2	13.25 feet	1.69 feet^2
3	13.97 feet	0.09 feet^2

Keep in mind that the values that can be observed for an individual's height (14 feet and 11 feet) do not change from population to population; what changes are the probabilities of seeing these values. Because of these differences in genotype frequencies, the mean and the variance are different for each population. The variance in the third population is very small since 99% of individuals in this population are of equal height, and thus the distance between those 99% of individuals' height (14 feet) and the mean height (13.97 feet) is very small.

Maximizing Variance

By understanding the components of variance (and having an intuitive understanding of what variance is), it is possible to describe the conditions that cause variance to be its maximum for a given experiment. Remember, variance is an expected value, and as such, it depends on two things: the values the random variable can take on and the probabilities of seeing those values. Obviously one way to increase variance is to increase the range between the smallest and the largest values the random variable can take on. For example, a random variable representing the toss of a fair six-sided die with the numbers 1, 3, 5, 7, 9, and 11 will have a larger variance than the usual six-sided die. But, what happens when you keep the values the random variable can take on constant, and just change the probabilities of seeing these values? What kind of probabilities cause variance to be its maximum in this situation? Remember, intuitively, variance is the expected difference between the values a random variable can take on and their mean, squared ($\sigma^2 = E[(X - \mu)^2]$). To make this as large as possible for fixed values of x, the probabilities (p) need to be such that the outcomes that are the furthest away from the mean (μ) have the highest probabilities of occurring. Variance is maximized when the two most extreme outcomes both have probability of 50% of occurring, and the probabilities of all the intermediate values are zero. In this case, only the most extreme values can occur, and the mean will be exactly centered between these two extremes (Figure 1.2). If instead the two extreme outcomes have unequal probabilities (and all other values have probability zero), the mean will be closer to the value with the higher probability. In this case the distances between the x values and the mean will be large for some x values, and small for others. However, the small distances will be more frequent than the large distances, so *on average* the squared distances between observations and the mean (and thus the variance) will be smaller. Figure 1.2 illustrates this.

Given this information, let's go back to the question we opened this chapter with: which of the two populations of beetles shown in Figure 1.1 displays the greatest biological variation? We said that to answer this question, we need to know how we measure variation. In Quantitative Genetics, we typically measure variation via variance. If we are interested in the number of spots per wing in these two populations, which population has the higher variance? The answer to this question is that population B has the higher variance in spot number. Thus, by this measure, population B has higher variation than population A, even though population A appears to display a larger level of biological diversity.



	number of spots per wing						
	1	2	3	4	5	μ	σ^2
p	0.2	0.2	0.2	0.2	0.2		
$x - \mu$	-2	-1	0	1	2	3	2
$p(x-\mu)^2$	0.8	0.2	0	0.2	0.8		
p	0.25	0.25	0	0.25	0.25		
$x - \mu$	-2	-1	0	1	2	3	2.5
$p(x-\mu)^2$	1	0.25	0	0.25	1		
p	0.5	0	0	0	0.5		
$x - \mu$	-2	-1	0	1	2	3	4
$p(x-\mu)^2$	2	0	0	0	2		
p	0.25	0	0	0	0.75		
$x - \mu$	-3	-2	-1	0	1	4	3
$p(x-\mu)^2$	2.25	0	0	0	0.75		

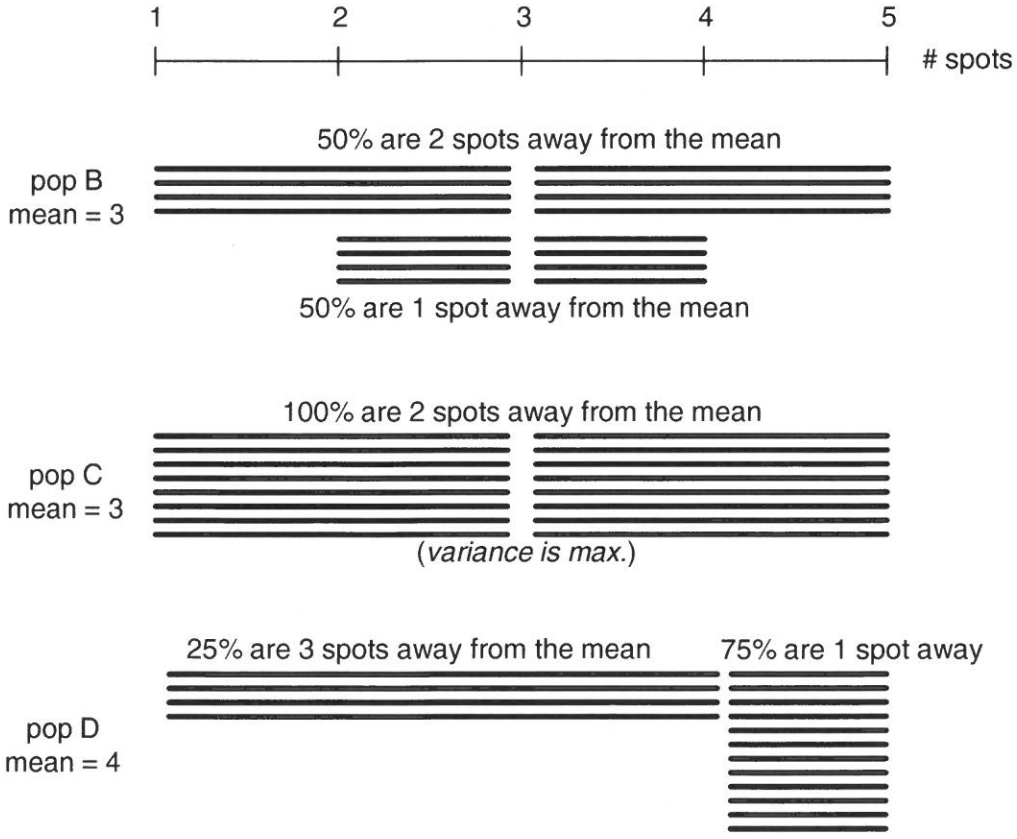


Figure 1.2 Variance is maximized when half the population is at one extreme (1 spot), and half at the other extreme (5 spots).