Quantitative traits and distributions

Lecture 1 ▪ Quantitative Genetics

Grum Gebreyesus ▪ ASAP-Bio Knowledge Hub

asap-bio.github.io/lec-qg-traits.html

Contents

1 Why quantitative genetics? 1

2 Metric traits: what we can and cannot observe 1

3 From one locus to many 2

4 Why phenotypes are normally distributed 2

5 The infinitesimal model 3

6 What the rest of the course covers 3

7 Exercises 4

1 Why quantitative genetics?

Genetics as most people first meet it is the genetics of **discrete traits**. A discrete trait takes a few distinct values, eye colour, the ability to roll the tongue, coat colour in Labradors. Such traits are usually controlled by the genotype at one or a few loci, and for many of them the responsible genes are known and understood. You can write down a genotype, apply Mendel's rules, and predict the phenotype.

Now consider milk yield, growth rate, feed intake, disease resistance or human height. None of these falls into neat categories. They vary continuously, and no single gene determines them.

DEFINITION

A **quantitative trait** is a trait that assumes values on a continuous scale. **Quantitative genetics** deals with the inheritance of variation in quantitative traits. Most quantitative traits are influenced by variation in many genes, each of small effect, together with the environment.

This is not a minor special case. Nearly every trait of economic importance in animal and plant breeding is quantitative. So we need a different language: one built on **expectations and probabilities** rather than on deterministic Mendelian ratios.

The strategy of this course is to model **one locus in complete detail**, and then extend that model to many loci. Almost everything that follows, breeding values, variance components, heritability, response to selection, is built on the one-locus foundation.

2 Metric traits: what we can and cannot observe

Quantitative genetics deals primarily with **metric traits**, traits we measure on a scale rather than classify into categories. Working with metric traits forces an important admission:

- The effect of any *single* gene cannot be observed directly.
- We can only speak about the trait in terms of **probability and statistics**.

This constraint determines the structure of the subject. Because the effect of an individual gene cannot be observed, the analysis cannot proceed gene by gene. It instead describes the *distribution* of the trait in a population and partitions the *variation* in that distribution into components attributable to genes and to environment. Probability and expectations, the subject of the next lecture, are therefore the working tools of the subject rather than a mathematical preliminary.

3 From one locus to many

Start with the model for a single locus. For a locus j with two alleles, the effect an individual carries is

$$g_j = (G_j - 2p_j)\,\alpha_j$$

where $G_j \in \{0, 1, 2\}$ counts copies of one allele, p_j is that allele's frequency and α_j is the allele substitution effect. The full derivation of this expression is the subject of the one-locus lecture; here we only need its shape.

Adding a second locus simply adds a second term:

$$y = \mu + g_1 + g_2 + e$$

and the general model for L loci is

$$y = \mu + \sum_{j=1}^{L} g_j + e \;=\; \mu + \sum_{j=1}^{L} (G_j - 2p_j)\,\alpha_j + e$$

where y is the observed phenotype, μ the overall mean, g_j the effect of the genotype at locus j, and e the residual or environmental effect. The individual's **breeding value** across all loci is $g = \sum_{j=1}^{L} g_j$.

EXPECTED VALUE OF AN OFFSPRING

Because a parent transmits one allele of each pair, the offspring's expected genetic value is the average of the parental breeding values:

$$E(y_{\text{off}}) = \mu + \sum_{j=1}^{L} \left(\tfrac{1}{2} g_{s,j} + \tfrac{1}{2} g_{d,j} \right)$$

with s the sire and d the dam. This relation underlies genetic improvement by selection, and reappears as the parent average in breeding-value prediction.

4 Why phenotypes are normally distributed

Height, milk yield, body weight and feed intake are all routinely assumed to be normally distributed. However, consider the building block:

$$g_j = (G_j - 2p_j)\alpha_j \quad \text{takes only three values, since } G_j \in \{0, 1, 2\}$$

A single gene effect is emphatically **not** normally distributed. It is a three-point discrete variable. So where does the familiar bell curve come from?

Derivation

1. A phenotype is a *sum* of many such three-point variables plus an environmental term: $y = \mu + \sum_{j=1}^{L} g_j + e$.
2. The **Central Limit Theorem** states that for independent random variables $X_1, \ldots, X_N$, $\dfrac{\sum_j X_j - E\left(\sum_j X_j\right)}{\sqrt{\mathrm{Var}\left(\sum_j X_j\right)}} \xrightarrow{D} N(0,1)$. Equivalently, for large N, $\sum_j X_j \xrightarrow{D} N\left(0, \mathrm{Var}\left(\sum_j X_j\right)\right)$.
3. Applying this to the genotype effects, with $E(g) = \sum_j E(g_j) = 0$ and $\mathrm{Var}(g) = \sigma_g^2$, the *sum* of a large enough number of individually non-normal gene effects is well approximated by a normal distribution, even though no single term is.

The normal density itself is

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \qquad -\infty < y < \infty$$

written $Y \sim N(\mu, \sigma^2)$.

INTERPRETATION

Normality of a quantitative trait is not an arbitrary convenience. It is a *consequence* of many small, independent genetic contributions adding up. This also tells you when to be careful: if a trait is governed by one or two large-effect genes, the normal approximation is poor and the distribution may be visibly multi-modal.

5 The infinitesimal model

Pushing the logic to its limit gives the model that underpins most of animal and plant breeding.

MAIN ASSUMPTIONS OF THE INFINITESIMAL MODEL

- Effects are **additive**.
- There is **no linkage disequilibrium** between loci.
- The phenotype y is unchanged whichever model we choose to describe it.

The third assumption is easily misread. The animal's actual phenotype does not change when a one-locus, two-locus or many-locus model is adopted. What changes is how the same total genetic effect is *distributed* among the terms:

$$\begin{aligned}
\text{One locus:} \quad & y = \mu + (G_1 - 2p_1)\alpha_1 + e \\
\text{Two loci:} \quad & y = \mu + (G_1 - 2p_1)\alpha_1 + (G_2 - 2p_2)\alpha_2 + e \\
\text{Infinitesimal:} \quad & y = \mu + \sum_{j=1}^{L} (G_j - 2p_j)\alpha_j + e
\end{aligned}$$

As more loci enter the model, the total genetic effect on y becomes more **diluted** among the individual α_j. Each locus explains less, but the sum explains the same. The infinitesimal model takes this to the extreme: infinitely many loci, each of infinitesimally small effect.

6 What the rest of the course covers

The topics ahead, in the order they build on one another:

1. Quantitative genetics of **one locus**
2. Quantitative genetics of **many loci**
3. **Genetic and environmental effects**, and their interactions
4. **Linkage disequilibrium** and gene mapping
5. Genetics of **multiple traits**
6. **Relationship and similarity** among relatives
7. Genetic change due to **selection** and **random events**

Two applied strands accompany these topics. **Artificial selection** aims to change the genetic properties of a population by preferentially breeding individuals whose genetics move the population in a desired direction. **Breeding value prediction** combines genetic, pedigree and phenotypic information into the best available prediction of which individuals carry the most desirable genetics, using linear models that pool information across many animals and several traits. **Breeding schemes** then organise this optimally, maximising genetic gain within physical, economic and biological constraints.

7 Exercises

These exercises use the R functions from the course (`QG_KU_functions.R`, `simY()`). The point is not the coding but the shape of the distributions you get.

Exercise 1.1

Simulate 100 phenotypes from a **single-locus** model and plot the density:

```
rm(list=ls(all=TRUE)); source("QG_KU_functions.R")
data <- simY(100, mu=-5)
plot.density(data$y)
```

Inspect the object with `str(data)`. What are the components `y`, `G`, `s` and `e`? Why is the density not bell-shaped?

`str(data)` returns a list of four: `y` the phenotypes, `G` the genotype codes (0/1/2) per locus, `s` the simulated allele substitution effects, and `e` the residuals.
With one locus the genetic contribution $g_1 = (G_1 - 2p_1)\alpha_1$ takes only three distinct values. The density is therefore a mixture of three narrow humps (one per genotype) smeared only by the environmental term e, not a single smooth bell.

Exercise 1.2

Repeat with an increasing number of loci:

```
data <- simY(100, n.loci=2, mu=0); plot.density(data$y)
data <- simY(100, n.loci=10, mu=0); plot.density(data$y)
```

```
data <- simY(100, n.loci=50, mu=0); plot.density(data$y)
data <- simY(100, n.loci=100, mu=0); plot.density(data$y)
```

At roughly how many loci does the distribution become visually indistinguishable from a normal curve? Relate what you see to the Central Limit Theorem.

With 2 loci you can still often see discrete humps (up to 9 genotype combinations). By about 10 loci the humps have merged; by 50 to 100 the density is smooth and symmetric and looks normal. This is the CLT in action: the sum $\sum_j g_j$ of many independent, individually non-normal three-point variables converges to a normal distribution. Note that the *number of loci needed* depends on how equal the effects are, a few large-effect loci among many small ones will keep the distribution skewed for longer.

Exercise 1.3

Conceptual, no computing. A colleague argues that because each gene effect g_j can only take three values, it is wrong to model a quantitative trait as normally distributed. What is the flaw in the argument, and under what circumstances would the colleague actually be right?

The flaw is confusing the distribution of a *single term* with the distribution of their *sum*. Normality is claimed for $y = \mu + \sum_j g_j + e$, not for any individual g_j.
The colleague is right, however, whenever the sum is dominated by a small number of large-effect loci. Then the CLT approximation fails and the phenotypic distribution can be visibly multimodal or skewed, coat colour scored on a numeric scale, or a trait with a major gene such as double-muscling, are practical examples.