

Livestock Breeding and Genomics

Peter von Rohr

2021-09-30

Contents

Preface	5
Prerequisites	5
Goals	5
Exam	6
Lecture Notes	6
1 Introduction	7
1.1 Terminology	7
1.2 History	8
1.3 Fundamental Questions	9
1.4 Improvement of Animal Populations	11
1.5 Statistics	12
1.6 Computer Science	12
2 Basics in Quantitative Genetics	15
2.1 Single Locus - Quantitative Trait	15
2.2 Frequencies	16
2.3 Hardy-Weinberg Equilibrium	17
2.4 Value and Mean	18
2.5 Variances	25
2.6 Extension To More Loci	27
2.7 Genetic Models	29
2.8 Appendix: Derivations	29
A Introduction To Linear Algebra	33
A.1 Glimpse Ahead	33
A.2 Vectors	33
A.3 Matrices	39
A.4 Systems Of Equations	42
A.5 Solving Systems of Linear Equations	44
B Introduction To R and RStudio	45

Preface

Livestock Breeding and Genomics is a masters course taught in the Institute of Agricultural Sciences at ETH Zurich. The aim of this course is to introduce the basic concepts of livestock breeding and genomics to the students.

Prerequisites

There are no explicit prerequisites for this course. All required concepts are explained during this course. Although basic knowledge in the following areas is definitely helpful for this course.

- **Linear Algebra:** Basic properties of and basic operations with vectors and matrices
- **Statistics:** Simple linear models and least squares
- **R:** Basics of how to work with data and how to do simple computations

Goals

The following goals are taken from the teaching system at ETH. The students are able to set up design matrices, the additive genetic relationship matrix and its inverse as well as the mixed model equations to estimate BLUP breeding values. The concept of BLUP breeding values are also extended to genomic selection. All concepts introduced during this course are not only presented during the lecture, but students are expected to work with the concepts on small example problems.

Specific Learning Goals

The students ...

- ... understand the basic concepts of livestock breeding and genomics.
- ... can separate the concepts of breeding and genomics from the techniques that are important in a livestock production enterprise.
- ... are familiar with the basic tools used in livestock breeding and genomics.
- ... know for which problems these tools can be applied.

- ... understand the genetic foundation behind the theory on which livestock breeding and genomics is based on.
- ... are able to extend the genetic theory into the statistical models used in livestock breeding and genomics.
- ... can extract the practical meaning of the results of the statistical analyses.

Exam

There will be a written exam during the lecture of the last week of the semester. The exam is going to be an open-book exam. All material that seems helpful to students will be allowed. The students must solve the exams on their own. The exam is not designed to be solved in teams.

Lecture Notes

These lecture notes contain all the relevant material for the exam. The appendix of the course notes contains two chapters, which introduce the necessary tools to work through the material of this course. Those tools consist of linear algebra and R. These chapters are mostly left to students to work through by themselves. Different students have different levels of skills in the topics of linear algebra and R. Hence not every student needs to go through the presented material at the same depth.

Chapter 1

Introduction

1.1 Terminology

The terms

- **Livestock Breeding** and
- **Animal Breeding**

are used interchangeably in this course. Although, the latter is a more general term, whereas the former focuses on livestock species, i.e. animals which are typically present on a farm such as cattle, pig, goat and sheep. Animal breeding in general could also include pets such as dogs or cats or even zoo animals. But the very interesting topic of breeding such species is outside of the scope of this lecture and is therefore not covered in these course notes.

The term *livestock breeding* is sometimes understood ambiguously. In general, most people do not differentiate between livestock breeding and animal husbandry or animal production. From a scientific point of view a *Livestock Breeder* is a person who owns a number of animals from which he or she selects parent animals and uses a designed mating scheme to achieve a certain goal with the offspring animals. Most commonly known are breeders of pet animals such as dogs, cats or birds which follow individual breeding goals which focus on specific phenotypic appearance or on special behavioral traits. Around the end of the 19th century, livestock breeders have realized that they have to collaborate in breeding associations to be able to effectively select parent animals from a large breeding population. This allowed them to achieve more robust selection responses in shorter amounts of time, especially for livestock species with long generation intervals such as cattle and horse. This development is nicely documented by the archive of Swiss agricultural history available at <https://www.histoiredurale.ch/afaahr/>. In summary, livestock breeders are primarily interested in selecting parent animals with the best genetic potential

to produce offspring animals that are closer to a breeding goal. From an economic point of view the sale of breeding animals to other farmers makes an important contribution to the economic result of the farm.

In contrast to livestock breeding, livestock production focuses on the aspects of keeping animals on farms to produce goods that can be sold. Nowadays, the goods produced by farm animals are mostly used in human nutrition. Due to the focus on the production aspects, the economic result of the production process is determined by the difference between monetary revenue for the products and the costs that are caused by the production of the goods.

Depending on the livestock species, the separation between breeding and production is more pronounced. In pigs, most farms are specialized into either breeding farms or production farms. Most cattle farms run as mixed operations which means that they are members of breeding organizations but they also run a production business. While the mixture of both operation types (breeding and production) on the same farm is not negative, it is conceptually important to separate breeding and production.

1.2 History

Livestock breeding and Genomics are two scientific areas which have experienced quite dramatic changes in the last few years. As already mentioned in the previous section, livestock breeding started over 100 years ago and is a much older discipline than Genomics.

1.2.1 Livestock Breeding

In principle, livestock breeding did exist for a very long time in a rather unsystematic form. Individual breeders always made choices about which animals they want to select as parents for the next generation of their livestock herds. Accounts that such early activities in livestock breeding happened as early as the Middle Ages are given in (Duerst, 1931) and (Arndorfer et al., 2010). But to the best of my knowledge it was only in the second half of the 20th century that the area of livestock breeding made some ground-breaking progress which spread all over the world. This progress was initiated by the researcher **Charles R. Henderson**. He and his team developed a solid methodology that is still applied up to the current day. The main achievement of Henderson and his team was to find a class of statistical models that are consistent with the theory of quantitative genetics described in (Falconer and Mackay, 1996) which is one reference among many other sources. Furthermore, the research groups lead by Henderson showed how to efficiently compute the results from the statistical models for large datasets.

1.2.2 Genomics

Genomics started with the Human Genome Project. The publication of the first draft of the complete Human genome (Venter et al., 2001) and a publication by (Meuwissen et al., 2001) that appeared in the same year made it possible to include information of complete genomes into statistical analyses. The process of including information from complete genomes into statistical analyses is what is understood by the term **genomics**.

The introduction of genomics methodologies in the area of livestock breeding caused a shift of paradigm. In large livestock breeding populations associations between certain genetic variants and the expression of desirable forms of phenotypic traits can be estimated using older breeding animals. The obtained estimation results can be used to assess the genetic potential of young animals which do not have any phenotypic observations available. This type of analysis is part of a procedure which is termed **genomic selection** and it allows for selecting breeding animals at a much younger age which shortens the generation interval.

The basic principle of how animals are selected as parents of future generations did not change, but the availability of different types of information and the amount of information that can be used to assess the genetic potential of a selection candidate changed dramatically since the invention of genomic technologies. Despite these rapid developments of new technologies, livestock breeders are still facing the following two fundamental questions.

1.3 Fundamental Questions

In livestock breeding and genomics, we are interested in addressing two fundamental questions that bothered breeders for a very long time. For this course, we put these two fundamental questions into the following form.

1. What is the *best* animal?
2. What can breeders do to obtain the *best* animal?

The term **best** is relative, because there is no **best** animal for all situations and all environments. Animals that show high performances in one environment, may not be able to produce as much in a different environment. One example for that might be Holstein cows in Europe or North America are able to produce a lot of milk, but they have difficulties to survive in Africa. Knowing that the environment plays an important role for livestock animals, we will be assuming that the animals that we are selecting, are more or less adapted to their environment.

Animals are usually described or characterized in terms of appearance or performance or a combination of both. In any case, we will be talking about **traits** where any trait is an observable or measurable characteristic of an animal. Examples of *observable* traits are

- coat color
- size
- muscling
- leg set
- udder conformation and many more.

Observable traits are mostly used to describe the appearance of an animal. In contrast to that, *measurable* traits are mostly used to describe the performance of an animal. Examples of measurable traits are

- body weight
- milk production
- protein and fat yield.

Note, it is important to distinguish between the observed or measured values of a trait which might be **red** coat color or 343 kg of body weight and the traits themselves which are just coat color or body weight. The observed or measured values of a trait are also called **phenotypes**.

1.3.1 Genotypes and Phenotypes

In livestock breeding we are mainly concerned with changing animal populations at the genetic level. The reason why we are interested in changing a population genetically is because parents do not pass their phenotypes to their offspring. Parents pass a random sample of their genes to their offspring. For each offspring every parent does transmit a different sample of their genes. From a genetic point of view, we want to know not only the most desirable phenotype, but also the most desirable **genotypes**. From the central dogma of molecular biology (https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology), it follows that an animal's genotype provides the genetic background of phenotypes. The relationship between phenotypes (P) and genotypes (G) can be summarized by the following equation (1.1)

$$P = G + E \tag{1.1}$$

where E represents the **environmental effects**. Because we want to change our populations at a genetic level, we are interested in the effect (G) of every genotype. In most cases, we are not able to directly observe or measure G . But we will see later on how we can estimate G based on measurements and observations of P and based on estimates of E . The estimates of G will later be called **breeding values** and those estimates will be used by breeders as information for their tools to improve animal populations. Those tools are being described in the following section.

1.4 Improvement of Animal Populations

The purpose of livestock breeding is to improve animal populations. Once an animal is conceived, the genotype is fixed¹ and cannot be improved anymore. Breeders can improve populations at the genetic level using the following two tools

1. selection
2. mating

1.4.1 Selection

Selection is the process to determine which individuals of a current population become parents of the next generation. The application of selection in a certain population over a certain time changes the animals in that population at the genetic level. The most familiar form of selection is **natural selection** which occurs in natural and wildlife populations. Natural selection is one of the great forces of evolution and it also affects domestic animals. All animals with lethal genetic defects are naturally selected against, i.e., they never become parents.

Although natural selection cannot be ignored for livestock species, what is most important for animal breeders is **artificial selection**. The idea behind artificial selection is simple. For a given trait all animals in a population are ranked according to their breeding value. From this list the animals ranking top are used as parents for the next generation. In most livestock populations, animal breeders are interested to improve their animals with respect to more than just one trait. When considering more than one trait, the question is how to come up with the ranking for the animals that are selected as potential parents. Several strategies to produce such a ranking based on a number of traits. It has been shown that using a weighted mean of the breeding values of all traits which is called **aggregate genotype** to rank all animals is an optimal procedure to be used as selection criterion (Hazel, 1943).

1.4.2 Mating

The second tool we have available as animal breeders is **mating**. In a mating scheme, we decide which of the selected male animals are bred to which selected female animals. There are a number of different rules that can be followed. The application of a given set of rules are summarized as mating system. There are three reasons for using a mating system.

1. producing offspring with extreme breeding values. When parents with extreme breeding values (high or low) are mated, offspring with extreme phenotypes can be expected. This is mostly used when a given trait is to be changed in one direction

¹Here we do not take into account new technologies such as gene editing.

2. make use of complementarity in parental traits. When neither of the parents is optimal, a mix of traits can be desirable. In such a case parental genotypes can be quite different. When parents of different breeds are mated, then this is called **crossbreeding**.
3. obtain positive effects due to heterosis. Hybrid vigor or heterosis in crossbreeding occurs when offspring performance exceeds the performance of the pure-breds.

There might also be other aspects that influence a mating system, e.g. to restrict the level of inbreeding or to consider optimum genetic contribution theory (Meuwissen and Sonesson, 1997).

1.5 Statistics

Several authors such as (Schaeffer, 2013) and (Gianola and Rosa, 2015) have reviewed the development of statistical methods in the area of animal breeding. Both authors mention that statistical methodology plays an important role in animal breeding. Most animal breeders are concerned with estimating or predicting breeding values. This is still done using a set of methods resulted from the theory developed by Charles Henderson and his team ((Henderson, 1953) and (Henderson, 1975)). These methods are known under the name of **BLUP**. BLUP shows some important regularization properties. These properties allow us to estimate or to predict many more unknown parameters than we have observations. In animal breeding, breeding values of all animals in a population can be predicted also for those animals for which we do not have observations. This is particularly important for traits which can only be observed in animals of one sex.

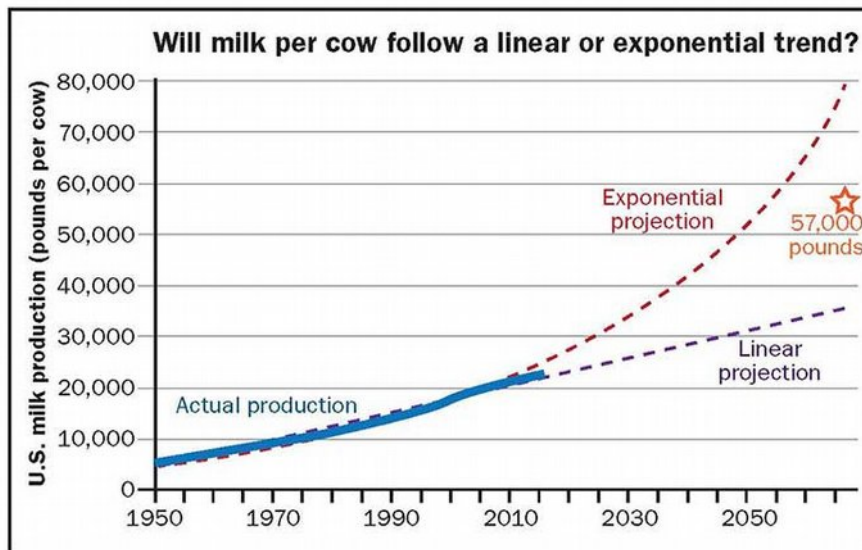
There are more methods with regularization properties. The so-called **Bayesian methods** are one example. Bayesian methods use the so-called Bayes theorem ((Bayes and Price, 1763) and https://en.wikipedia.org/wiki/Bayes%27_theorem) to come up with parameter estimates. Although Bayesian methods are much older than other methods such as BLUP, they were only introduced into practical animals breeding in the early 1990's. Important pioneering papers for the use of Bayesian methods in animal breeding are (Gianola and Foulley, 1982) and (Gianola and Fernando, 1986). The reasons for the late adoption of Bayesian methods are certainly related to development of cheap computing infrastructure. This is described in more detail in subsection 1.6.

1.6 Computer Science

The development of computing power is often summarized by the so-called **Moore's Law** ((Moore, 1965) and https://en.wikipedia.org/wiki/Moore%27s_law). Moore's law is not a law in the sense of mathematics or physics, but it

is a prediction that Gordon Moore² made as early as 1965. He predicted that the number of components that could be placed on a certain integrated circuit would double roughly every 18 months between 1959 and 1975. This prediction was generalized into a statement that the general computing performance could be doubled every 18 months. In retro-spect this was more or less true for the last 50 years. This considerable increase in computing performance had also a dramatic impact in the costs of a certain computation.

When comparing the development of computing performance with the performance of livestock animals, there is an obvious relation between the two. This means the performance increase of livestock animals was in part facilitated by the development of cheap computing power. The two figures 1.1 and 1.2 compare the two developments. The first diagram shows the annual milk production per cow.



Milk Performance per Cow

(Source: <https://hoards.com/article-20808-what-will-dairy-cows-and-farms-look-like-in-50-years.html>)

Figure 1.1: Yearly Milk Yield per Cow in the USA

The Figure 1.2 below shows the development of computing power according to Moore's law.

²One of the co-founders and a director of Intel

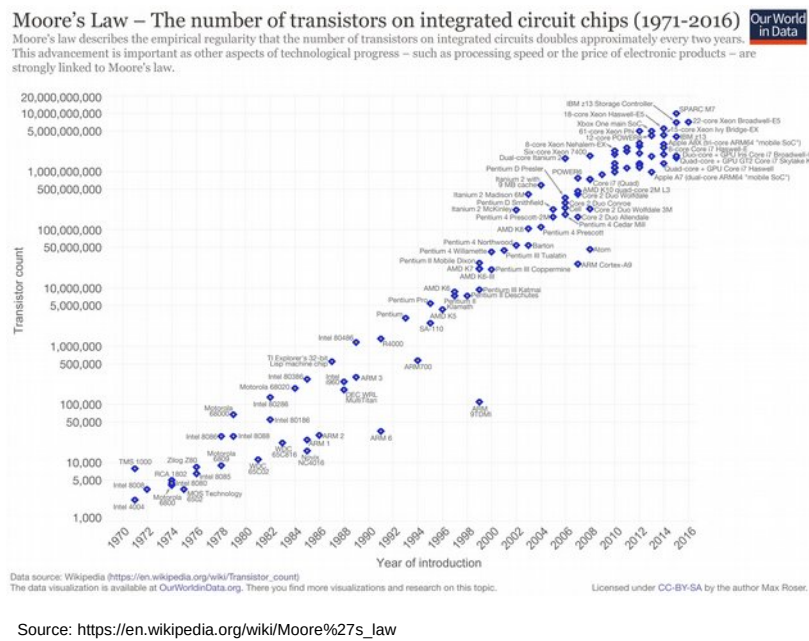


Figure 1.2: Computing Performance According To Moore's Law

Chapter 2

Basics in Quantitative Genetics

As already mentioned in section 1.3.1, the central dogma of molecular biology tells us that the genotype is the basics of any phenotypic expression. The genotype of an individual is composed of a number of genes which are also called **loci**. In this section, we start with the simplest possible genetic architecture where the genotype is composed by just one locus. The connection between the genotype and the phenotype is modeled according to equation (1.1). The phenotype is assumed to be a quantitative trait. That means we are not looking at binary or categorical traits. Categorical traits can just take a limited number of different levels. Examples of categorical traits are the horn status in cattle or certain color characteristics. Quantitative traits do not take discrete levels but they show specific distributions.

2.1 Single Locus - Quantitative Trait

In Livestock there are not many examples where a quantitative trait is influenced by just one locus. But this case helps in understanding the foundation of more complex genetic architectures. We start by looking at the following idealized population (Figure 2.1).

2.1.1 Terminology

The different genetic variants that are present at our Locus G are called **alleles**. When looking at all individuals in the population for our locus, we have two different alleles G_1 and G_2 . Hence, we call the locus G to be a **bi-allelic** locus. In any given individual of the population, the two alleles of the locus G together are called the individuals **genotype**. All possible combinations of the two alleles

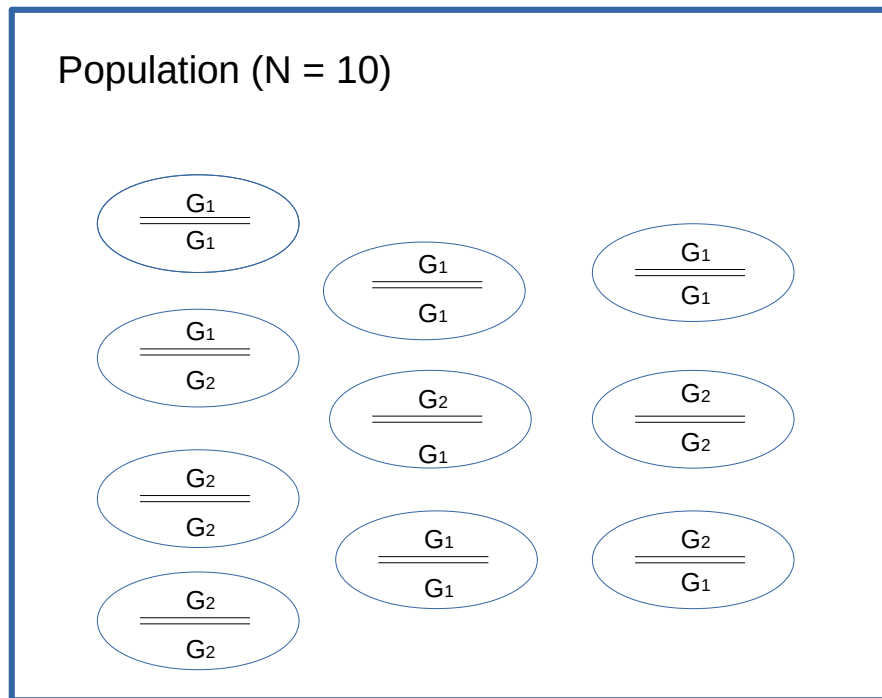


Figure 2.1: Idealized Population With A Single Locus

at the locus G leads to a total number of three genotypes. It is important to mention that the order of the alleles in a given genotype is not important. Hence, G_1G_2 and G_2G_1 are the same genotype. The two genotypes G_1G_1 and G_2G_2 are called **homozygous** and the genotype G_1G_2 is called **heterozygous**.

2.2 Frequencies

To be able to characterize our population with respect to the locus of interest, we are first looking at some frequencies. These are measures of how often a certain allele or genotype does occur in our population. For our example population shown in Figure 2.1, the **genotype frequencies** are

$$\begin{aligned}
f(G_1G_1) &= \frac{4}{10} = 0.4 \\
f(G_1G_2) &= \frac{3}{10} = 0.3 \\
f(G_2G_2) &= \frac{3}{10} = 0.3
\end{aligned} \tag{2.1}$$

The **allele frequencies** can be determined either by counting or they can be computed from the genotype frequencies.

$$\begin{aligned}
f(G_1) &= f(G_1G_1) + \frac{1}{2} * f(G_1G_2) = 0.55 \\
f(G_2) &= f(G_2G_2) + \frac{1}{2} * f(G_1G_2) = 0.45
\end{aligned} \tag{2.2}$$

2.3 Hardy-Weinberg Equilibrium

The Hardy-Weinberg equilibrium is the central law of how allele frequencies and genotype frequencies are related in an idealized population. Given the allele frequencies

$$\begin{aligned}
f(G_1) &= p \\
f(G_2) &= q = 1 - p
\end{aligned} \tag{2.3}$$

During mating, we assume that in an idealized population alleles are combined independently. This leads to the genotype frequencies shown in Table 2.1.

Summing up the heterozygous frequencies leads to

$$\begin{aligned}
f(G_1G_1) &= p^2 \\
f(G_1G_2) &= 2pq \\
f(G_2G_2) &= q^2
\end{aligned} \tag{2.4}$$

Table 2.1: Genotype Frequencies under Hardy-Weinberg equilibrium

Alleles	G_1	G_2
G_1	$f(G_1G_1) = p^2$	$f(G_1G_2) = p * q$
G_2	$f(G_1G_2) = p * q$	$f(G_2G_2) = q^2$

Comparing these expected genotype frequencies in a idealized population under the Hardy-Weinberg equilibrium to what we found for the small example population in Figure 2.1, we can clearly say that the small example population is not in Hardy-Weinberg equilibrium.

2.4 Value and Mean

Our goal is still to improve our population at the genetic level. The term improvement implies the need for a quantitative assessment of our trait of interest. Furthermore, we have to be able to associate the genotypes in the population to the quantitative values of our trait.

2.4.1 Genotypic Values

The values V_{ij} to each genotype G_iG_j are assigned as shown in Figure 2.2.

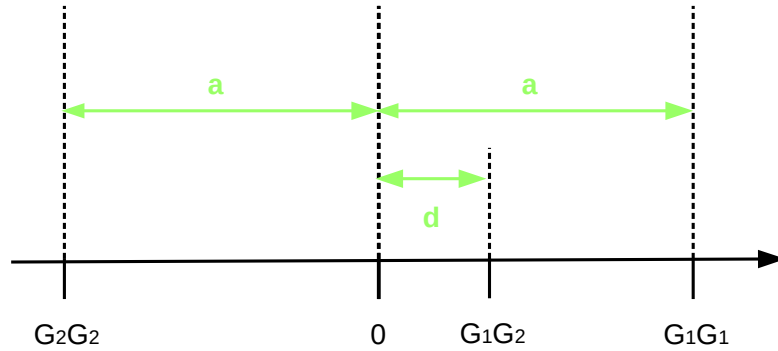


Figure 2.2: Genotypic Values

The origin of the genotypic values is placed in the middle between the two homozygous genotypes G_2G_2 and G_1G_1 . Here we are assuming that G_1 is the favorable allele. This leads to values of $+a$ for genotype G_1G_1 and of $-a$ for genotype G_2G_2 . The value of genotype G_1G_2 is set to d and is called dominance deviation. Table 2.2 summarizes the values for all genotypes.

Table 2.2: Values for all Genotypes

Variable	Genotype	Values
V_{11}	G_1G_1	a
V_{12}	G_1G_2	d
V_{22}	G_2G_2	-a

2.4.2 Population Mean

For the complete population, we can compute the **population mean** (μ) of all values at the locus G . This mean corresponds to the expected value and is computed as

$$\begin{aligned}
 \mu &= V_{11} * f(G_1G_1) + V_{12} * f(G_1G_2) + V_{22} * f(G_2G_2) \\
 &= a * p^2 + d * 2pq + (-a) * q^2 \\
 &= (p - q)a + 2pqd
 \end{aligned} \tag{2.5}$$

The population mean depends on the values a and d and on the allele frequencies p and q . The larger the difference between p and q the more influence the value a has in μ , because for very different p and q the product $2pq$ is very small. On the other hand, if $p = q = 0.5$, then $\mu = 0.5d$. For loci with $d = 0$, the population mean $\mu = (p - q)a$ and hence, if in addition we have $p = q$, then $\mu = 0$.

2.4.3 Breeding Values

The term **breeding value** is defined as shown in Definition 2.1.

Definition 2.1 (Breeding Value). The breeding value of an animal i is defined as two times the difference between the mean value of offsprings of animal i and the population mean.

Applying this definition and using the parameters that we have computed so far leads to the following formulas for the breeding value of an animal with a certain genotype.

2.4.3.1 Breeding value for G_1G_1

Assume that we have a given parent S with a genotype G_1G_1 and we want to compute its breeding value. Let us further suppose that our single parent S is mated to a potentially infinite number of animals from the idealized population, then we can deduce the following mean genotypic value for the offspring of parent S .

	Mates of S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_1) = 1$	$f(G_1G_1) = p$	$f(G_1G_2) = q$

Because parent S has genotype G_1G_1 , the frequency $f(G_1)$ of a G_1 allele coming from S is 1 and the frequency $f(G_2)$ of a G_2 allele is 0. The expected genetic value (μ_{11}) of the offspring of animal S can be computed as

$$\mu_{11} = p * a + q * d \quad (2.6)$$

Applying definition 2.1, we can compute the breeding value (BV_{11}) for animal S as shown in equation (2.7) while using the results given by equations (2.6) and (2.5).

$$\begin{aligned}
BV_{11} &= 2 * (\mu_{11} - \mu) \\
&= 2 (pa + qd - [(p - q)a + 2pqd]) \\
&= 2 (pa + qd - (p - q)a - 2pqd) \\
&= 2 (qd + qa - 2pqd) \\
&= 2 (qa + qd(1 - 2p)) \\
&= 2q (a + d(1 - 2p)) \\
&= 2q (a + (q - p)d)
\end{aligned} \quad (2.7)$$

Breeding values for parents with genotypes G_2G_2 and G_1G_2 are derived analogously.

2.4.3.2 Breeding value for G_2G_2

First, we determine the expected genotypic value for offsprings of a parent S with genotype G_2G_2

	Mates of parent S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_2) = 1$	$f(G_1G_2) = p$	$f(G_2G_2) = q$

The expected genetic value (μ_{22}) of the offspring of animal S can be computed as

$$\mu_{22} = pd - qa \quad (2.8)$$

The breeding value BV_{22} corresponds to

$$\begin{aligned}
 BV_{22} &= 2 * (\mu_{22} - \mu) \\
 &= 2 (pd - qa - [(p - q)a + 2pqd]) \\
 &= 2 (pd - qa - (p - q)a - 2pqd) \\
 &= 2 (pd - pa - 2pqd) \\
 &= 2 (-pa + p(1 - 2q)d) \\
 &= -2p (a + (q - p)d)
 \end{aligned} \quad (2.9)$$

2.4.3.3 Breeding value for G_1G_2

The genotype frequencies of the offsprings of a parent S with a genotype G_1G_2 is determined in the following table.

	Mates of parent S	
	$f(G_1) = p$	$f(G_2) = q$
Parent S		
$f(G_1) = 0.5$	$f(G_1G_1) = 0.5p$	$f(G_1G_2) = 0.5q$
$f(G_2) = 0.5$	$f(G_1G_2) = 0.5p$	$f(G_2G_2) = 0.5q$

The expected mean genotypic value of the offsprings of parent S with genotype G_1G_2 is computed as

$$\mu_{12} = 0.5pa + 0.5d - 0.5qa = 0.5 [(p - q)a + d] \quad (2.10)$$

The breeding value BV_{12} corresponds to

$$\begin{aligned}
ZW_{12} &= 2 * (\mu_{12} - \mu) \\
&= 2 (0.5(p - q)a + 0.5d - [(p - q)a + 2pqd]) \\
&= 2 (0.5pa - 0.5qa + 0.5d - pa + qa - 2pqd) \\
&= 2 (0.5(q - p)a + (0.5 - 2pq)d) \\
&= (q - p)a + (1 - 4pq)d \\
&= (q - p)a + (p^2 + 2pq + q^2 - 4pq)d \\
&= (q - p)a + (p^2 - 2pq + q^2)d \\
&= (q - p)a + (q - p)^2d \\
&= (q - p) [a + (q - p)d]
\end{aligned} \tag{2.11}$$

2.4.3.4 Summary of Breeding Values

The term $a + (q - p)d$ appears in all three breeding values. We replace this term by α and summarize the results in the following table.

Genotype	Breeding Value
G_1G_1	$2q\alpha$
G_1G_2	$(q - p)\alpha$
G_2G_2	$-2p\alpha$

2.4.4 Allele Substitution

Comparing the genotype G_2G_2 with the genotype G_1G_2 , one of the differences is in the number of G_1 -alleles. G_2G_2 has zero G_1 -alleles and G_1G_2 has one G_1 -allele.

Let us imagine that we take animal i with a G_2G_2 genotype and use the CRISPR-CAS genome editing technology to replace one of the G_2 alleles in animal i by a G_1 allele (see Figure 2.3). After applying the gene editing procedure to animal i at locus G , animal i would have genotype G_1G_2 .

Due to the application of genome editing at locus G of animal i the breeding value changed. Before the genome editing procedure it was BV_{22} and after genome editing the breeding value of animal i corresponds to BV_{12} . So the effect of replacing a G_2 allele by a G_1 allele on the breeding value corresponds to the difference $BV_{12} - BV_{22}$. The computation of this difference between the breeding value BV_{12} and BV_{22} results in

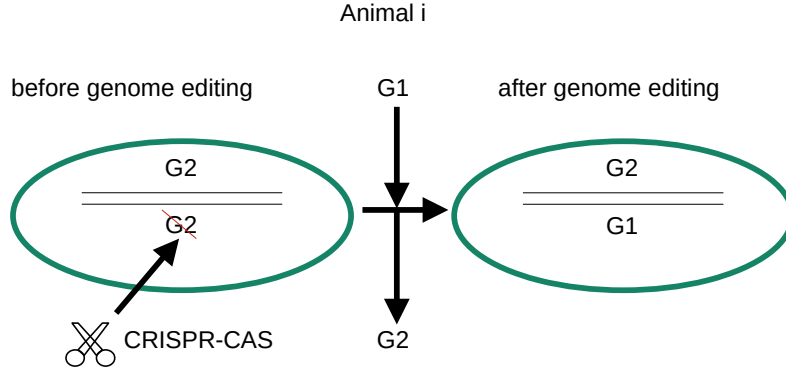


Figure 2.3: Schematic Depiction of Genome Editing on Animal i

$$\begin{aligned}
 BV_{12} - BV_{22} &= (q - p)\alpha - (-2p\alpha) \\
 &= (q - p)\alpha + 2p\alpha \\
 &= (q - p + 2p)\alpha \\
 &= (q + p)\alpha \\
 &= \alpha
 \end{aligned} \tag{2.12}$$

The analogous computation can be done by comparing the breeding values BV_{11} and BV_{12} .

$$\begin{aligned}
 BV_{11} - BV_{12} &= 2q\alpha - (q - p)\alpha \\
 &= (2q - (q - p))\alpha \\
 &= \alpha
 \end{aligned} \tag{2.13}$$

Because the differences between breeding values computed in (2.12) and (2.13) are equal, we can conclude that the breeding values show a linear dependence on the number of G_1 alleles. This is the reason why the breeding values are also called additive effects, because adding a further G_1 allele instead of a G_2 allele has always the same effect on the breeding values, namely just adding the constant allele substitution effect α .

2.4.5 Dominance Deviation

When looking at the difference between the genotypic value V_{ij} and the breeding value BV_{ij} for each of the three genotypes, we get the following results.

$$\begin{aligned}
 V_{11} - BV_{11} &= a - 2q\alpha \\
 &= a - 2q[a + (q - p)d] \\
 &= a - 2qa - 2q(q - p)d \\
 &= a(1 - 2q) - 2q^2d + 2pqd \\
 &= [(p - q)a + 2pqd] - 2q^2d \\
 &= \mu + D_{11}
 \end{aligned} \tag{2.14}$$

$$\begin{aligned}
 V_{12} - BV_{12} &= d - (q - p)\alpha \\
 &= d - (q - p)[a + (q - p)d] \\
 &= [(p - q)a + 2pqd] + 2pqd \\
 &= \mu + D_{12}
 \end{aligned} \tag{2.15}$$

$$\begin{aligned}
 V_{22} - BV_{22} &= -a - (-2p\alpha) \\
 &= -a + 2p[a + (q - p)d] \\
 &= [(p - q)a + 2pqd] - 2p^2d \\
 &= \mu + D_{22}
 \end{aligned}$$

The difference all contain the population mean μ plus a certain deviation. This deviation term is called **dominance deviation**.

2.4.6 Summary of Values

The following table summarizes all genotypic values all breeding values and the dominance deviations.

Genotyp G_iG_j	genotypic value V_{ij}	Breeding Value BV_{ij}	Dominance Deviation D_{ij}
G_1G_1	a	$2q\alpha$	$-2q^2d$
G_1G_2	d	$(q - p)\alpha$	$2pqd$
G_2G_2	$-a$	$-2p\alpha$	$-2p^2d$

The formulas in the above shown table assume that G_1 is the favorable allele with frequency $f(G_1) = p$. The allele frequency of G_2 is $f(G_2) = q$. Since we have a bi-allelic locus $p + q = 1$.

Based on the definition of dominance deviation, the genotypic values V_{ij} can be decomposed into the components population mean (μ), breeding value (BV_{ij}) and dominance deviation (D_{ij}) according to equation (2.16).

$$V_{ij} = \mu + BV_{ij} + D_{ij} \quad (2.16)$$

Taking expected values on both sides of equation (2.16) and knowing that the population mean μ was defined as the expected value of the genotypic values in the population, i.e. $E[V] = \mu$, it follows that the expected values of both the breeding values and the dominance deviations must be 0. More formally, we have

$$\begin{aligned} E[V] &= E[\mu + BV + D] \\ &= E[\mu] + E[BV] + E[D] \\ &= \mu \end{aligned} \quad (2.17)$$

From the last line in equation (2.17), it follows that $E[BV] = E[D] = 0$. This also shows that both breeding values and dominance deviations are defined as deviation from a given mean.

2.5 Variances

The population mean μ and derived from that the breeding values were defined as expected values. Their main purpose is to assess the state of a given population with respect to a certain genetic locus and its effect on a phenotypic trait of interest. One of our primary goals in livestock breeding is to improve the populations at the genetic level through the means of selection and mating. Selection of potential parents that produce offspring that are closer to our breeding goals is only possible, if the selection candidates show a certain level of variation in the traits that we are interested in. In populations where there is no variation which means that all individuals are exactly at the same level, it is not possible to select potential parents for the next generation.

In statistics the measure that is most often used to assess variation in a certain population is called **variance**. For any given discrete random variable X the variance is defined as the second central moment of X which is computed as shown in equation (2.18).

$$Var[X] = \sum_{x_i \in \mathcal{X}} (x_i - \mu_X)^2 * f(x_i) \quad (2.18)$$

where \mathcal{X} : set of all possible x -values
 $f(x_i)$ probability that x assumes the value of x_i
 μ_X expected value $E[X]$ of X

In this section we will be focusing on separating the obtained variances into different components according to their causative sources. Applying the definition of variance given in equation (2.18) to the genotypic values V_{ij} , we obtain the following expression.

$$\begin{aligned} \sigma_G^2 = Var[V] &= (V_{11} - \mu)^2 * f(G_1G_1) \\ &+ (V_{12} - \mu)^2 * f(G_1G_2) \\ &+ (V_{22} - \mu)^2 * f(G_2G_2) \end{aligned} \quad (2.19)$$

where $\mu = (p - q)a + 2pqd$ the population mean.

Based on the decomposition of the genotypic value V_{ij} given in (2.16), the difference between V_{ij} and μ can be written as the sum of the breeding value and the dominance deviation. Then σ_G^2 can be written as

$$\begin{aligned} \sigma_G^2 = Var[V] &= (BV_{11} + D_{11})^2 * f(G_1G_1) \\ &+ (BV_{12} + D_{12})^2 * f(G_1G_2) \\ &+ (BV_{22} + D_{22})^2 * f(G_2G_2) \end{aligned} \quad (2.20)$$

Inserting the expressions for the breeding values BV_{ij} and for the dominance deviation D_{ij} found earlier and simplifying the equation leads to the result in (2.21). A more detailed derivation of σ_G^2 is given in the appendix (2.8) of this chapter.

$$\begin{aligned} \sigma_G^2 &= 2pq\alpha^2 + (2pqd)^2 \\ &= \sigma_A^2 + \sigma_D^2 \end{aligned} \quad (2.21)$$

The formula in equation (2.21) shows that σ_G^2 consists of two components. The first component σ_A^2 is called the **genetic additive variance** and the second component σ_D^2 is termed **dominance variance**. As shown in equation (2.27) σ_A^2 corresponds to the variance of the breeding values. Because we have already

seen that the breeding values are additive in the number of favorable alleles, σ_A^2 is called genetic additive variance. Because σ_D^2 corresponds to the variance of the dominance deviation effects (see equation (2.29)) it is called dominance variance.

2.6 Extension To More Loci

When only a single locus is considered, the genotypic values (V_{ij}) can be decomposed according to equation (2.16) into population mean, breeding value and dominance deviation. When a genotype refers to more than one locus, the genotypic value may contain an additional deviation caused by non-additive combination effects.

2.6.1 Epistatic Interaction

Let V_A be the genotypic value of locus A and V_B denote the genotypic value of a second locus B , then the total genotypic value V attributed to both loci A and B can be written as

$$V = V_A + V_B + I_{AB} \quad (2.22)$$

where I_{AB} is the deviation from additive combination of these genotypic values. When computing the population mean earlier in this chapter, we assumed that I was zero for all combinations of genotypes. If I is not zero for any combination of genes at different loci, those genes are said to **interact** with each other or to exhibit **epistasis**. The deviation I is called interaction deviation or epistatic deviation. If I is zero, the genes are called to act additively between loci. Hence *additive action* may mean different things. When referring to one locus, it means absence of dominance. When referring to different loci, it means absence of epistasis.

Interaction between loci may occur between pairs or between higher numbers of different loci. The complex nature of higher order interactions, i.e., interactions between higher number of loci does not need to concern us. Because in the total genotypic value V , interaction deviations of all sorts are treated together in an overall interaction deviation I .

Applying the decomposition of the genotypic values V_A of locus A and V_B of locus B as shown in (2.16) leads to

$$\begin{aligned} V &= V_A + V_B + I_{AB} \\ &= \mu_A + BV_A + D_A + \mu_B + BV_B + D_B + I_{AB} \end{aligned} \quad (2.23)$$

Collecting terms in (2.23) as follows

$$\begin{aligned}
 \mu &= \mu_A + \mu_B \\
 U &= BV_A + BV_B \\
 D &= D_A + D_B \\
 I &= IAB
 \end{aligned} \tag{2.24}$$

The decomposition shown in (2.23) and the collection of variables (see (2.24)) can be generalized to more than two loci. This leads to the following generalized form of decomposing the overall total genotype V for the case of multiple loci affecting a certain trait of interest.

$$V = \mu + U + D + I \tag{2.25}$$

where U is the sum of the breeding values attributable to the separate loci and D is the sum of all dominance deviations. For our purposes in livestock breeding where we want to assess the genetic potential of a selection candidate to be a parent of offspring forming the next generation, the **breeding value** is the most important quantity. The breeding value is of primary importance because a given parent passes a random sample of its alleles to its offspring. We have seen in section 2.4.4 that breeding values are additive in the number of favorable alleles. Hence the more favorable alleles a given parent passes to its offspring the higher the breeding value of this parent.

On the other hand, the dominance deviation measures the effect of a certain genotype occurring in an individual and the interaction deviation estimates the effects of combining different genotypes at different loci in the genome. But because parents do not pass complete genotypes nor do they pass stretches of DNA with several loci, but only a random collection of its alleles, it is really the breeding value that is of primary importance in assessing the genetic potential of a given selection candidate.

2.6.2 Interaction Variance

If genotypes at different loci show epistatic interaction effects as described in section 2.6.1, the interactions give rise to an additional variance component called V_I , which is the variance of interaction deviations. This new variance component V_I can be further decomposed into sub-components. The first level of sub-components is according to the number of loci that are considered. Two-way interactions involve two loci, three-way interactions consider three loci and in general n -way interactions arise from n different loci. The next level of subdivision is according to whether they include additive effects, dominance deviations or both.

In general it can be said that for practical purposes, interaction effects explain only a very small amount of the overall variation. As already mentioned in section @ref(#epistatic-interaction) for livestock breeding, we are mostly interested in the additive effects. This is also true when looking at the variance components. Hence dominance variance and variances of interaction deviations are not used very often in practical livestock breeding application.

2.7 Genetic Models

In this chapter, we have seen how to model the genetic basis of a quantitative trait when a single locus affects the trait of interest. We call this a single-locus model. When several loci have an effect on a certain trait, then we talk about a **polygenic model**. Letting the number of loci affecting a certain phenotype tend to infinity, the resulting model is called **infinitesimal model**.

From a statistical point of view, the breeding values in an infinitesimal model are considered as a random effect with a known distribution. Due to the central limit theorem, this distribution is assumed to be a normal distribution. The central limit theorem says that the distribution of any sum of a large number of very small effects converges to a normal distribution. For our case where a given trait of interest is thought to be influenced by a large number of genetic loci each having a small effect, the sum of the breeding values of all loci together can be approximated by a normal distribution. Figure (2.4) shows the distribution for a sum of 10, 100 and 1000 components each. The histograms show a better approximation to the normal distribution the larger the number of components considered in the sum.

2.7.1 Model Usage In Routine Evaluations

Traditional prediction of breeding values before the introduction of genomic selection is based on the infinitesimal model. When genomic selection was introduced which takes into account the information from a large number of loci, genomic breeding values are estimated using a polygenic model.

2.8 Appendix: Derivations

This section shows how the genetic variance in equation (2.21) is computed.

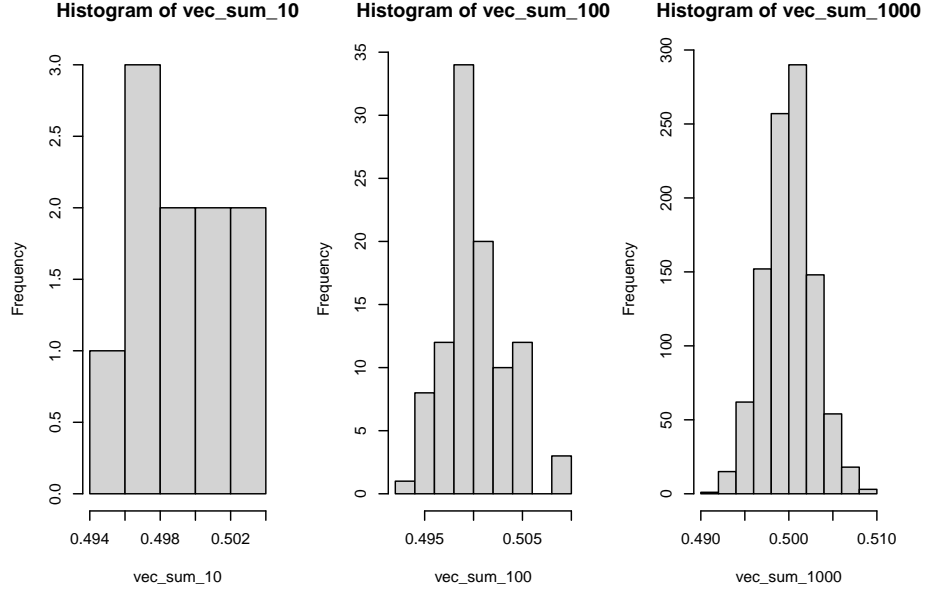


Figure 2.4: Distribution of Sums of Different Numbers of Components

$$\begin{aligned}
\sigma_G^2 &= (BV_{11} + D_{11})^2 * p^2 \\
&+ (BV_{12} + D_{12})^2 * 2pq \\
&+ (BV_{22} + D_{22})^2 * q^2 \\
&= (2q\alpha - 2q^2d)^2 * p^2 \\
&+ ((q - p)\alpha + 2pqd)^2 * 2pq \\
&+ (-2p\alpha - 2p^2d)^2 * q^2 \\
&= (4q^2\alpha^2 - 8q^3d\alpha + 4q^4d^2) * p^2 \\
&+ (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2 - 4(q - p)pqd\alpha + 4p^2q^2d^2) * 2pq \\
&+ (4p^2\alpha^2 + 8p^3d\alpha + 4p^4\alpha^2) * q^2 \\
&= 4p^2q^2\alpha^2 - 8p^2q^3d\alpha + 4p^2q^4d^2 \\
&+ 2pq^3\alpha^2 - 4p^2q^2\alpha^2 + 2p^3q\alpha^2 \\
&- 8p^3q^2d\alpha + 8p^2q^3d\alpha + 8p^3q^3d^2 \\
&+ 4p^2q^2\alpha^2 + 8p^3q^2d\alpha + 4p^4q^2d^2 \\
&= 4p^2q^2\alpha^2 + 4p^2q^4d^2 \\
&+ 2pq^3\alpha^2 + 2p^3q\alpha^2 \\
&+ 8p^3q^3d^2 \\
&+ 4p^4q^2d^2 \\
&= 2pq\alpha^2 (p^2 + 2pq + q^2) \\
&+ (2pqd)^2 (p^2 + 2pq + q^2) \\
&= 2pq\alpha^2 + (2pqd)^2 \\
&= \sigma_A^2 + \sigma_D^2
\end{aligned} \tag{2.26}$$

From the last two lines of (2.26) it follows that $\sigma_A^2 = 2pq\alpha^2$ and $\sigma_D^2 = (2pqd)^2$. It can be shown that σ_A^2 corresponds to the squared breeding values times the associated genotype frequencies. Because the expected values of the breeding values is zero, σ_A^2 is equivalent to the variance of the breeding values.

$$\begin{aligned}
\sigma_A^2 &= \text{Var}[BV] = (BV_{11} - E[BV])^2 * f(G_1G_1) \\
&\quad + (BV_{12} - E[BV])^2 * f(G_1G_2) \\
&\quad + (BV_{22} - E[BV])^2 * f(G_2G_2) \\
&= BV_{11}^2 * f(G_1G_1) + BV_{12}^2 * f(G_1G_2) + BV_{22}^2 * f(G_2G_2) \\
&= (2q\alpha)^2 * p^2 + ((q-p)\alpha)^2 * 2pq + (-2p\alpha)^2 * q^2 \\
&= 4p^2q^2\alpha^2 + (q^2\alpha^2 - 2pq\alpha^2 + p^2\alpha^2) * 2pq + 4p^2q^2\alpha^2 \\
&= 8p^2q^2\alpha^2 + 2pq^3\alpha^2 - 4p^2q^2\alpha^2 + 2p^3q\alpha^2 \\
&= 4p^2q^2\alpha^2 + 2pq^3\alpha^2 + 2p^3q\alpha^2 \\
&= 2pq\alpha^2 (2pq + q^2 + p^2) \\
&= 2pq\alpha^2
\end{aligned} \tag{2.27}$$

In the above derivation in (2.27) of the variance of the breeding values, we were using the fact that the expected value $E[BV] = 0$. This can be shown more formally as follows

$$\begin{aligned}
E[BV] &= BV_{11} * f(G_1G_1) + BV_{12} * f(G_1G_2) + BV_{22} * f(G_2G_2) \\
&= 2q\alpha * p^2 + (q-p)\alpha * 2pq + (-2p\alpha) * q^2 \\
&= 2p^2q\alpha + 2pq^2\alpha - 2p^2q\alpha - 2pq^2\alpha \\
&= 0
\end{aligned} \tag{2.28}$$

Similarly to (2.27) we can show that σ_D^2 corresponds to the squared dominance deviations times the frequencies of the corresponding genotypes. That is the reason why σ_D^2 is called dominance variance.

$$\begin{aligned}
\sigma_D^2 &= D_{11}^2 * f(G_1G_1) + D_{12}^2 * f(G_1G_2) + D_{22}^2 * f(G_2G_2) \\
&= (-2q^2d)^2 * p^2 + (2pqd)^2 * 2pq + (-2p^2d)^2 * q^2 \\
&= 4p^2q^4d^2 + 8p^3q^3d^2 + 4p^4q^2d^2 \\
&= 4p^2q^2d^2 (q^2 + 2pq + p^2) \\
&= 4p^2q^2d^2
\end{aligned} \tag{2.29}$$

Appendix A

Introduction To Linear Algebra

Linear Algebra is a large area. A short overview can be seen in a video (<https://youtu.be/kjBOesZCoqc>) by (3Blue1Brown, 2016a). In this course we only need the following three topics from the large field of linear algebra.

1. Vectors
2. Matrices and
3. Systems of linear equations.

A.1 Glimpse Ahead

The central topic of this course is the prediction of breeding values. Most approaches to predict breeding values require the solution of large systems of linear equations. These systems of equations are written down using vectors and matrices. Hence the three mentioned topics are important to understand at a level that they can be used as tools for the prediction of breeding values.

A.2 Vectors

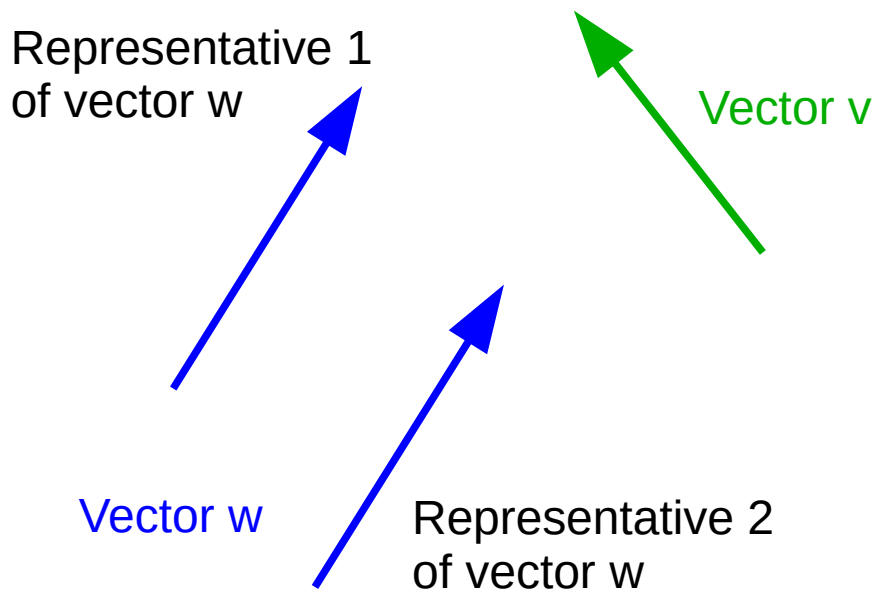
The material of this section is largely based on the video tutorial (https://youtu.be/fNk_zzaMoSs) from (3Blue1Brown, 2016b). We try to give a summarized transcript of the video. The vector is the fundamental building block of linear algebra. There are three different but related concepts about what vectors are. We call them

1. the physics perspective
2. the computer science perspective and
3. the mathematics perspective.

The mathematics perspective tries to provide a very general concept, saying that anything can be a vector as long as, one can add two vectors or a vector can be multiplied by a factor and the result of both operations is a vector again. For what we want to use vectors for in the context of livestock breeding and genomics, the mathematics perspective is not so useful, hence we ignore it from now on.

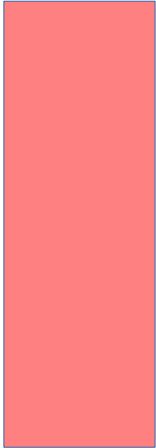
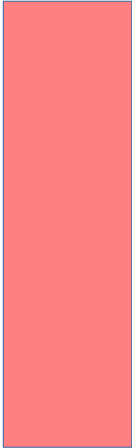
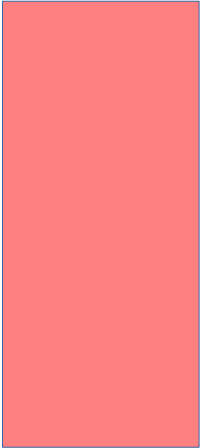
A.2.1 Physics Perspective

The physics perspective is that vectors are arrows with a certain **length** and a **direction** they are pointing to. As long as length and direction are the same, the arrows can be moved around and they are still the same vector. Different arrows with the same length and the same direction are called **representatives** of the same vector. Vectors that are in a flat plane are called two-dimensional. Those who are sitting in the same Euclidean space that we are all living in, are called three-dimensional.



A.2.2 Computer Science Perspective

In the computer science perspective vectors are ordered list of numbers. Later we will see that vectors can also contain more general objects like strings. As an example, we assume that we are analyzing carcasses and the only thing we know about a carcass is its slaughter-weight (SW) and its price (P). The different carcasses can then be represented by a pair of numbers the first being the slaughter-weight and the second being the price. It is important to note here, that the order of the number matters. In terms of vectors, here each carcass is represented by a two-dimensional vector.

	Carcass 1	Carcass 2	Carcass 3
			
Weight	290 kg	265 kg	320 kg
Price	1943 Fr	1829 Fr	1984 Fr

A.2.3 Geometric Context

Some basic properties of vectors are introduced using the geometric context, that a vector is an arrow located in a certain coordinate system with its tail sitting at the origin of the coordinate system. This is a little bit different from the physics perspective (see A.2.1) where the arrow can sit anywhere in space.

In linear algebra it is almost always the case that vectors are rooted at the origin. Once we understand the properties of vectors in the context of arrows in space, we can then translate these properties to the list-of-numbers point of view (see A.2.2) considering the coordinates of the vectors.

A.2.4 Coordinate System

It is important to introduce the coordinate system, because this will be the basis of the correspondence between the two perspectives of linear algebra. For the moment, we focus on two dimensions. The horizontal line is called the x-axis and the vertical line is called the y-axis. The place where the two lines intersect is called the origin. An arbitrary length is chosen to represent 1. The coordinates of a vector is a pair of numbers that give instructions for how to get from the tail of that vector at the origin to its tip. The first number tells you how far to walk along the x-axis (positive numbers indicating rightward motion, negative numbers indicating leftward motion) and the second number tell you how far to walk parallel to the y-axis (positive numbers indicating upward motion, negative numbers indicating downward motion).

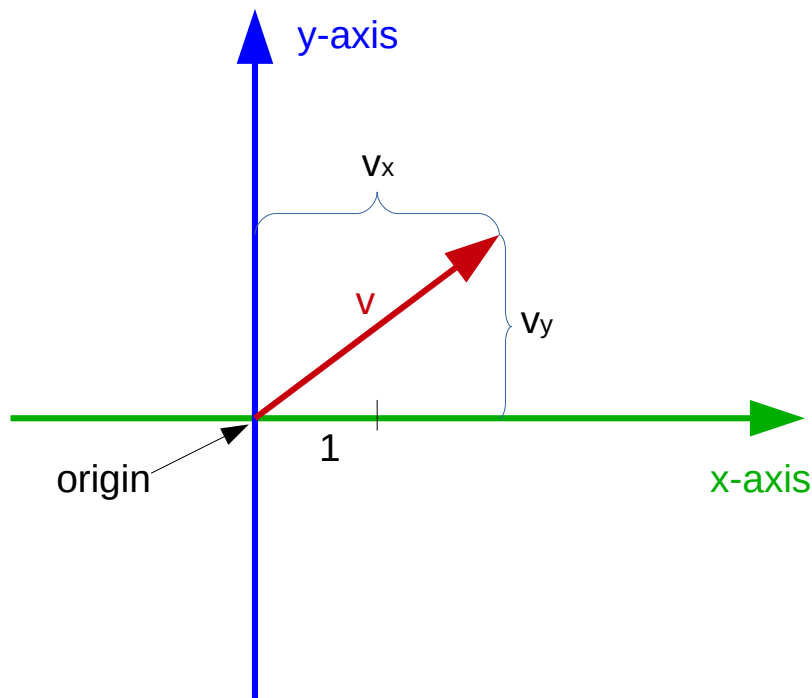


Figure A.1: Coordinate System

A.2.5 Vector Operations

The vectors by themselves can be pretty interesting objects, but they get really useful when considering some operations that we can perform on them. Here we consider three basic operations.

1. addition
2. multiplication by a scalar number and
3. dot product

A.2.5.1 Addition

Let us assume, we have two vectors v and w . To add these two vectors, move the second one such that its tail sits at the tip of the first one. Then draw a new vector from the tail of the first one to the tip of the second one. The new vector corresponds to the sum of the two vectors (Figure A.2).

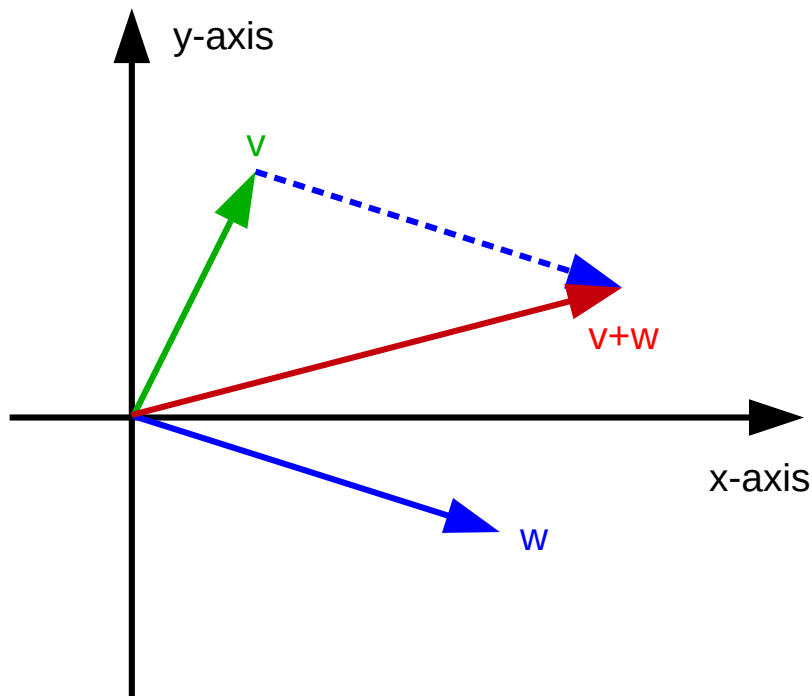


Figure A.2: Addition of two vectors

Numerically, vector addition corresponds to summing up each of the coordinates individually. Hence if we have two vectors v and w with their coordinates given as

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix}, w = \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

then the sum $v + w$ has coordinates

$$v + w = \begin{bmatrix} v_x + w_x \\ v_y + w_y \end{bmatrix}$$

A.2.5.2 Multiplication by a Scalar Number

This operation is best understood by looking at a few examples. If we take the number 2 and multiply it by a certain vector v , this means that we stretch out the vector v such that it is 2 times as long as the original vector. Multiplication of a vector with positive numbers does not change the direction of the vector. Multiplying a vector v with a negative number like -0.5 then the direction gets flipped around and then squished by 0.5.

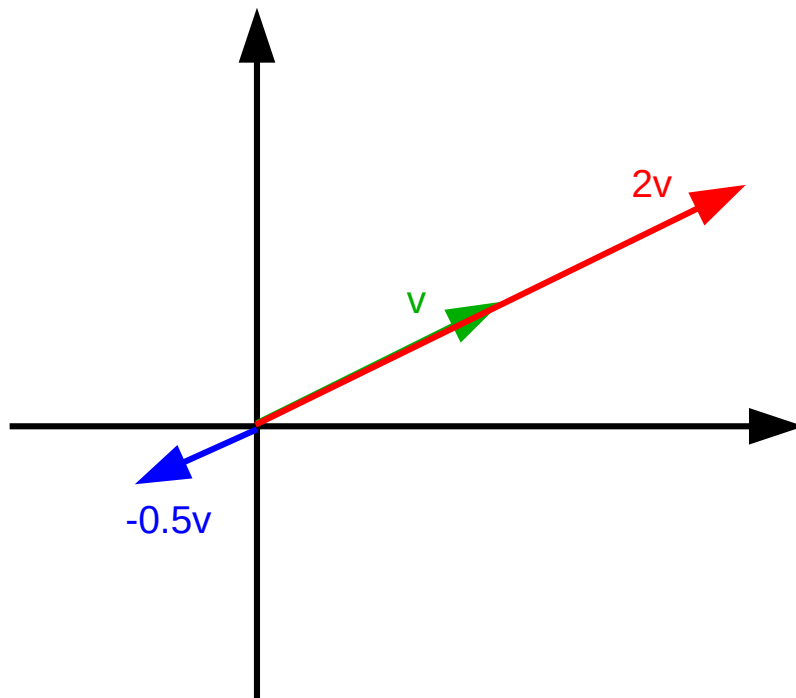


Figure A.3: Scalar Multiplication of Vectors

The operation of multiplying a vector by a given number, like 2 or -0.5 is also

called **scaling** and that is the reason why in linear algebra the numbers like 2 and -0.5 are called **scalar** numbers or just scalars. Numerically, stretching a vector by a given number like 2, corresponds to multiplying each of the coordinate components by that factor 2. For a vector v with coordinate components v_x and v_y , the vector $2v$ has coordinates $2v_x$ and $2v_y$

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix}, \quad 2v = \begin{bmatrix} 2v_x \\ 2v_y \end{bmatrix}$$

A.2.5.3 Dot Product

The dot product is explained in a different video that can be seen on <https://youtu.be/LyGKycYT2v0>. Numerically, if you have two vectors of the same dimension, meaning two lists of numbers of the same length, e.g. v and w then their dot product $v \cdot w$ can be computed by pairing up all of the coordinates, multiplying these pairs together and adding the result. So the vectors

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} w_x \\ w_y \end{bmatrix}$$

their dot product $v \cdot w$ then is computed as

$$v \cdot w = v_x * w_x + v_y * w_y$$

A.3 Matrices

The introduction to the topic of matrices is available from <https://youtu.be/kYB8IZa5AuE> and <https://youtu.be/XkY2DOUCWMU>. An $m \times n$ matrix is a table-like object of $m * n$ numbers arranged in m rows and n columns. In general the $m \times n$ matrix A has the following structure.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

The $m*n$ numbers inside of the square brackets are called elements of the matrix. The element of matrix A that is in row i and in column j is called a_{ij} or $(A)_{ij}$. As an example

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 5 & 1 & 2 \end{bmatrix}$$

is a 2×3 matrix. In the first row the second element corresponds to $(A)_{12} = a_{12} = 3$. An $n \times n$ matrix (i.e. a matrix with equal numbers of rows and columns)

is called a **quadratic** matrix. Two matrices A and B are called **equal**, if they have the same number of rows and columns and if the corresponding elements are the same, i.e.

$$(A)_{ij} = (B)_{ij} \text{ for all } i \text{ and } j$$

A.3.1 Special Matrices

The following matrices are special and are used in special cases.

- **Nullmatrix:** The $m \times n$ matrix 0 is called Nullmatrix, if each element is equal to zero.
- **Upper Triangular Matrix:** The square matrix R is called upper triangular matrix, if $(R)_{ij} = 0$ for $i > j$.
- **Lower Triangular Matrix:** The square matrix L is called lower triangular matrix, if $(L)_{ij} = 0$ for $i < j$.
- **Diagonal Matrix:** The square matrix D is called diagonal matrix, if $(D)_{ij} = 0$ for $i \neq j$.
- **Identity Matrix:** The diagonal matrix I is called identity matrix, if all diagonal elements $(I)_{ii} = 1$.
- **Column Vector:** A $m \times 1$ matrix is often called a column vector.
- **Row Vector:** A $1 \times n$ matrix is often called a row vector.

A.3.2 Matrix Operations

The following operations with matrices are defined.

A.3.2.1 Addition

For two $m \times n$ matrices A and B , their sum $A + B$ is again a $m \times n$ matrix with each element corresponding to the sum of the corresponding elements from A and B . Hence, we can write

$$(A + B)_{ij} = (A)_{ij} + (B)_{ij} \text{ for all } i \text{ and } j$$

A.3.2.2 Multiplication with a Number

A $m \times n$ matrix A is multiplied by a number α by multiplying every element $(A)_{ij}$ of A with α . The result $\alpha * A$ is computed as $(\alpha * A)_{ij} = \alpha * (A)_{ij}$ for all i and j .

A.3.2.3 Multiplication of two Matrices

Given a $m \times n$ matrix A and a $n \times p$ matrix B , their matrix product AB is a $m \times p$ matrix with

$$(AB)_{ij} = \sum_{k=1}^n (A)_{ik} * (B)_{kj} = (A)_{i1} * (B)_{1j} + (A)_{i2} * (B)_{2j} + \dots + (A)_{in} * (B)_{nj}$$

A.3.2.4 Laws of Matrix Operations

- **Commutativity:** For two $m \times n$ matrices A and B the addition is commutative, i.e. $A + B = B + A$.
- **Associativity of addition:** For $m \times n$ matrices A , B and C , the addition is associative, i.e., $A + (B + C) = (A + B) + C$
- **Associativity of multiplication:** For a $m \times n$ matrix A , a $n \times p$ matrix B and a $p \times q$ matrix C , the multiplication is associative, i.e., $A(BC) = (AB)C$
- **Distributivity:** For $m \times n$ matrices A and B and $n \times p$ matrices C and D , the distributive law holds, i.e., $(A + B)C = AC + BC$ and $A(C + D) = AC + AD$

A.3.2.5 Matrix Transpose

Given a $m \times n$ matrix A , then the $n \times m$ matrix A^T is called its **transpose**, if $(A^T)_{ij} = A_{ji}$. The matrix A is called **symmetric**, if $A = A^T$. For every matrix A the transpose of the transpose is the matrix itself, i.e., $(A^T)^T = A$. For any $m \times n$ matrices A and B , the transpose $(A + B)^T$ of their sum $(A + B)$ is computed as

$$(A + B)^T = A^T + B^T$$

For every $m \times n$ matrix A and every $n \times p$ matrix B , it holds that

$$(AB)^T = B^T A^T$$

A.3.2.6 Inverse of a Matrix

In this section, we are looking at square matrices. The **inverse** X of a square matrix A is defined as the square matrix that satisfies the condition $AX = I$. If the inverse matrix X exists, then the matrix A is called invertable. If X does not exist, A is called singular. If the inverse of a matrix A exists, it is uniquely determined and we call it A^{-1} .

Let us assume two invertable $n \times n$ matrices A and B , then the following equations hold

1. $A^{-1}A = I$
2. A^{-1} is invertable and $(A^{-1})^{-1} = A$
3. I is invertable and $I^{-1} = I$

4. AB is invertable and $(AB)^{-1} = B^{-1}A^{-1}$
5. A^T is invertable and $(A^T)^{-1} = (A^{-1})^T$

For every square matrix A , the following statements are equivalent.

1. A is invertable
2. The system of equations $Ax = b$ is solvable for every b .
3. The system of equations $Ax = 0$ has only the trivial solution $x = 0$.

A.3.2.7 Orthogonal Matrices

A square matrix A is called **orthogonal**, if the condition $A^T A = I$ holds. For two orthogonal matrices A and B , the following statements hold.

1. A is invertable and $A^{-1} = A^T$
2. A^{-1} is orthogonal
3. AB is orthogonal
4. I is orthogonal

A.4 Systems Of Equations

Systems of linear equations are introduced based on (Nipp and Stoffer, 2002) and (Searle, 1971). Solving systems of linear equations is one of the fundamental tasks of linear algebra. We start with a general example of a system of linear equations which is given as

$$\begin{aligned} x_1 + 2x_2 &= 5 \\ 2x_1 + 3x_2 &= 8 \end{aligned} \tag{A.1}$$

In (A.1) we are given a system of linear equations with two equations and two unknowns x_1 and x_2 . The aim is to find numeric values for x_1 and x_2 such that both equations are satisfied. Inserting the values $x_1 = 1$ and $x_2 = 2$ into the above equations show that they are both satisfied. Hence the set $L = \{x_1 = 1, x_2 = 2\}$ consisting of the values for x_1 and x_2 that satisfy both equations is called a solution or a solution set for the above shown equations.

In general, a linear system of equations consists of m equations and n unknowns. In the example (A.1), $m = 2$ and $n = 2$.

The example in (A.2) does not have any solutions.

$$\begin{aligned} x_1 + x_2 &= 4 \\ 2x_1 + 2x_2 &= 5 \end{aligned} \tag{A.2}$$

This can be seen, that if the first equation in (A.2) is multiplied by 2, we get $2x_1 + 2x_2 = 8$ which contradicts the second equation shown in (A.2).

A system with $m = 2$ equations and $n = 3$ unknowns is shown in (A.3).

$$\begin{aligned}x_1 - x_2 + x_3 &= 2 \\ 2x_1 + x_2 - x_3 &= 4\end{aligned}\tag{A.3}$$

There are infinitely many solutions consisting of $x_1 = 2$, $x_2 = \alpha$ and $x_3 = \alpha$ for any real number α .

The examples in (A.1), (A.2) and (A.3) already show all possible cases that may occur when solving linear systems of equations. The question is how to determine the set of all solutions of a system of linear equations.

A.4.1 Matrix-Vector Notation

So far, we have written systems of linear equations explicitly in the sense that every equation was written on one line. For small systems this is not a problem. But when the number of equations (m) and the number of unknowns (n) get very large, the explicit notation is no longer feasible. Hence, we need a notation that can also be used for large systems of equations. The so-called matrix-vector notation provides an efficient way to write down large systems of equations very efficiently.

We return to the example given by (A.1) and we define the matrix A to be

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix},$$

the vector x to be

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

and the vector y to be

$$y = \begin{bmatrix} 5 \\ 8 \end{bmatrix},$$

With these definitions, we can write the system of equations given in (A.1) using matrix-vector notation as

$$A \cdot x = y\tag{A.4}$$

A.5 Solving Systems of Linear Equations

If matrix A in (A.4) is not singular, i.e. the inverse Matrix A^{-1} of A does exist, the solution x to (A.4) can be written as $x = A^{-1}y$. This result is obtained by pre-multiplying both sides of (A.4) with A^{-1} and since a matrix times its inverse results in the identity matrix I , the solution is obtained as

$$\begin{aligned} A \cdot x &= y \\ A^{-1} \cdot A \cdot x &= A^{-1} \cdot y \\ I \cdot x &= A^{-1} \cdot y \\ x &= A^{-1} \cdot y \end{aligned} \tag{A.5}$$

For systems of equations with a singular matrix A , solutions can be found, if the equations are **consistent**. The linear equations $Ax = y$ are consistent, if any linear relationship existing among the rows of A also exist among the corresponding elements of y . As a simple example, the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 21 \end{bmatrix}$$

are consistent. In the matrix on the left the second row corresponds to three times the first row and in the vector on the right, the second element is also three times the first element. In contrast the equations

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 24 \end{bmatrix}$$

are not consistent. From this example, we can already see that non-consistent equations do not have any solutions. But consistent equations $Ax = y$ have a solution which can be written as $x = Gy$ if and only if, $AGA = A$ which means that G is a so-called generalized inverse of A . The matrix G is often written as A^{-} . The proof of this statement is given on page 9 of (Searle, 1971).

Appendix B

Introduction To R and RStudio

According to the website <https://www.r-project.org/>, R (R Core Team, 2018) is a free software environment for statistical computing and graphics. Because R is free and it is available for the most commonly used operating systems such as Windows, MacOSX and Linux, it has become very popular in statistics and in data science. Furthermore, R can be extended with user-contributed code and documentation (called R-packages) in a very easy and standardised way. The number of available R-packages is growing rapidly and has reached

RStudio (<https://www.rstudio.com/>) is a private company that among a large number of different products distributes the RStudio Integrated Development Environment (IDE) for R. A great number of different resources about R and RStudio IDE is available. Some of them are listed here.

- **R:** From the R website, there is an introduction to R available at: <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>.
- **RStudio IDE:** The link <https://resources.rstudio.com/> points to a series of webinars introducing R and RStudio.

Bibliography

3Blue1Brown (2016a). Essence of linear algebra preview.

3Blue1Brown (2016b). What are vectors.

Arndorfer, M., Blauensteiner, P., Berger, B., Falschlunger, G., Fischerleitner, F., Frohofer, W., Gaber, R., Holzner, W., Kajtna, B., Koller, B., Kriechbaum, M., Machatschek, M., Neunteufel, M., Spitzenberger, F., Michael, S., Stekovics, E., Vogl, C., Vogl-Lukasser, B., Wallner, R. M., Wilfing, A., and Zipser, P. (2010). *Wie viele Arten braucht der Mensch?* Böhlau, Wien, Köln, Weimar.

Bayes, T. and Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton. *Philosophical Transactions of the Royal Society of London*, 53:370—418.

Duerst, J. U. (1931). *Grundlagen der Rinderzucht*. Springer, Berlin; Heidelberg.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Addison Wesley Longman Limited, Essex, 4 edition.

Gianola, D. and Fernando, R. L. (1986). Bayesian Methods in Animal Breeding Theory. *Journal of Animal Science*, 63(1):217–244.

Gianola, D. and Foulley, J. L. (1982). Non linear prediction of latent genetic liability with binary expression: An empirical Bayes approach. *Proc. 2nd World Congr. Genet. Appl. Livest. Prod., Madrid, Spain*, 7:293—303.

Gianola, D. and Rosa, G. J. M. (2015). One Hundred Years of Statistical Developments in Animal Breeding. *Annual Review of Animal Biosciences*, 3:13.1—13.38.

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28(6):476–490.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226—252.

- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423—449.
- Meuwissen, T. H. and Sonesson, A. K. (1997). Maximizing the Response of Selection with a Predefined Rate of Inbreeding: Overlapping Generations. *Journal of Animal Science*, 76(10):2575–2583.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157:1819–1829.
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38.
- Nipp, K. and Stoffer, D. (2002). *Lineare Algebra*. vdf Hochschulverlag, Zurich, 5 edition.
- R Core Team (2018). R: A language and environment for statistical computing.
- Schaeffer, L. R. (2013). *History of Genetic Evaluation Methods in Dairy Cattle*. self.
- Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, New York, wiley clas edition.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., Mckusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., ru Ji, R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., Mccawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M.,

Rodriguez, R., hui Rogers, Y., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., ó, R. G., Campbell, M. J., Blick, L., Caminha, M., Carnes-stine, J., Caulk, P., hui Chiang, Y., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., Mcdaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291(5507):1–38.