Means and Variances

Gustavo de los Campos (gustavoc@msu.edu)

September 8th, 2025

1 Single-locus model

This section presents the math underlying a single-locus model. For a quantitative phenotype we can represent a phenotypic measurement (Y) as the sum of a genetic value (G) and an environmental effect (E)

$$Y = G + E$$

where G is the expected value of the phenotype given the genotype of the individual; therefore, $[E] = 0$.

In a bi-allelic locus with allleles $\{A_1; A_2\}$ we can have three genotypes; therefore, G can take on three values

- $E[Y|A_1A_1] = G(A_1A_1)$,
- $E[Y|A_1A_2] = G(A_1A_2)$, and
- $E[Y|A_2A_2] = G(A_2A_2)$.

We will study the statistical/quantitative genetics properties of this model (the mean, the variance and the borad sense heritability) using two equivalent approaches. First, we will derive the mean and variance under Hardy-Weinberg assumptions (HW). Under these assumptions we can predict the proportion of each of the genotypes (A_1A_1, A_1A_2, A_2A_2) from allele frequencies, once we have these probabilities, we can derive the expected value and variance of phenotype using standard statistical rules. Second, we will derive the same parameters (additive and dominance effects, and the average effect of an allele substitution) using a regression approach.

1.1 Means and Variance under HWE

Consider a single locus with two alleles (A_1/A_2) and let p denote the frequency of allele A_1, that is $P(A_1) = p$, implying, $P(A_2) = 1 - p = q$.

For this locus we have three possible genotypes (A_1A_1, A_1A_2, A_2A_2), Table 1 gives the expected genotype frequencies under Hardy-Weinberg Equilibrium (HWE).

Table 1: Genotypes and expected frequencies under Hardy-Weinberg Equilibrium

Genotype	HWE Frequency
A_1A_1	p^2
A_1A_2	$2pq$
A_2A_2	q^2

Genetic values

Recall that the genetic value of an individual (G) is the expected phenotype given the genotype. As noted eaerlier, since we have three possible genotypes we can have three genetic values (or means):

- $G(A_1A_1)$,
- $G(A_1A_2)$, and
- $G(A_2A_2)$.

Following the parameterization used in Falconer & Mackay (1996), for the single locus model we represent the three possible genetic values using

- $G(A_1A_1) = \mu + a$,
- $G(A_1A_2) = \mu + d$, and
- $G(A_2A_2) = \mu - a$.

These three means can be simplified by subtracting μ such that

- $G(A_1A_1) = a$,
- $G(A_1A_2) = d$, and
- $G(A_2A_2) = -a$.

With this parameterizations, we model the three means (all represented as deviations from μ) using two parameters (a, and d); this is represented graphically in Figure 1 which includes four cases: additive model ($d = 0$), partial dominance ($|d| < |a|$), complete dominance ($|a| = |d|$), and over-dominance ($|d| > |a|$).

Tecnicnally, a is one-half of the difference between the genetic value of the two homozygous, $a = [G(A_1A_1) - G(A_2A_2)]/2$, the sign of a depends on the allele chosen as reference. On the other hand, d is the deviation from the additive model needed to fit the genetic value of the heterozygous: $d = G(A_1A_2) - [G(A_1A_1) - a]$

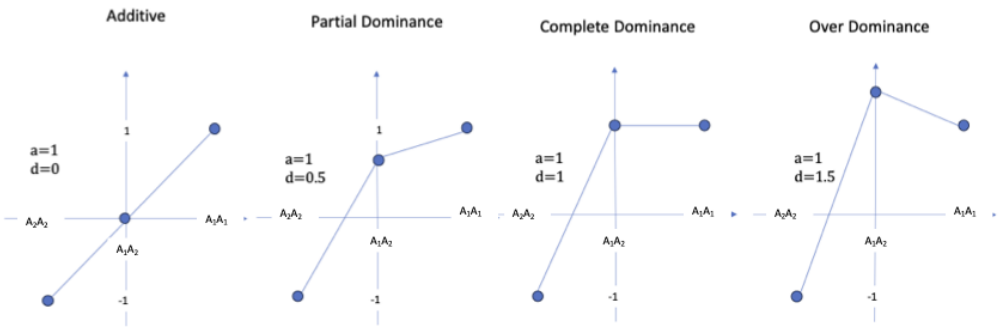


Figure 1: Figure 2: Means in the Single Locus Model

The population mean under Hardy-Weinberg equilibrium

To compute the population mean, we update **Table 1** adding a column for the genetic values. After some simplifications, including using $p^2 - q^2 = p - q$, we get $E(G) = a(p - q) + 2pqd$.

Table 2: Computing the population mean

Genotype	HWE Frequency	G	Freq×G
A_1A_1	p^2	a	p^2a
A_1A_2	$2pq$	d	$2pqd$
A_2A_2	q^2	-a	$-q^2a$
Expected value			$a(p - q) + 2pqd$

Genotype	HWE Frequency	G	Freq×G

Additive action: A model with $d = 0$ is a strictly additive model in which the genetic value of the heterozygous is half-way in between the genetic value of the two homozygous.

Genetic variance in the single locus model under Hardy-Weinberg Equilibirum

Recall that $\sigma^2_G = Var(G) = E(G^2) - E(G)^2$. We already have an expression for $E(G)$; to derive $E(G^2)$, we can update Table 3 by adding a column for G^2, multiplying by frequency, and taking column-wise sums to obtain the expected values.

Table 3: Computing the Genetic Variance

Genotype	HWE Frequency	G	$Freq \times G$	$Freq \times G^2$
A_1A_1	p^2	a	p^2a	p^2a^2
A_1A_2	$2pq$	d	$2pqd$	$2pqd^2$
A_2A_2	q^2	-a	$-q^2a$	q^2a^2
Expected Value			$a(p-q)+2pqd$	$a^2(p^2+q^2)+2pqd^2$

Using $E[G^2] = a^2(p^2 + q^2) + 2pqd^2$ and $E(G) = a(p - q) + 2pqd$ in $Var(G) = E(G^2) - E(G)^2$ leads, after simplifications, to the following expression for the genetic variance of a single locus:

$\sigma^2_G = 2pq(a + d[q - p])^2 + (2pqd)^2$.

Broad-sense heritability (in a single locus model)

The broad sense heritability of a trait is the proportion of the phenotypic variance explained by genetic factors. Consider a model with

$$Y = G + E$$

where $G = E[Y|Genotype|$ is a genetic values and E is an environmental effect. The phenotypic variance can be decomposed as follows $Var(Y) = Var(G) + Var(E) + 2Cov(G, E)$. If we assume that genetic and environmental factors are un-correlated, we have

$$Var(Y) = Var(G) + Var(E)$$

In this setting, the proportion of the phenotypic variance explained by genetic vactors (or broad sense heritability) is

$$H^2 = \frac{Var(G)}{Var(G) + Var(E)}$$

From the derivations presented above we have an expression mapping from allele frequencies and effects ($\{a, d\}$) into $Var(G)$.

Departures from HWE

Above we derive the mean, vairance, and heritability under HWE. The same framework can be applied in absence of HWE by repacing the predicted probabilities of each of the genotypes with the observed ones, i.e., replacing in the tables presented above with p^2 with $p(A_1A_1)$, q^2 with $p(A_2A_2)$ and $2pq$ with $p(A_1A_2)$.

1.2 Regression approach

Here, we will recast the model presented above as a linear regression problem.

Let $X \in \{0, 1, 2\}$ be a variable counting the number of copies of A_1 alleles and $Z = X - 1$ such that

- $Z = -1$ if the genotype is A_2A_2,
- $Z = 0$ if the genotype is A_1A_2, and
- $Z = 1$ if the phenotype is A_1A_1.

Now consider a dummy variable for the heterozygous genotype $H = \{1, \, if \; heterozygous \; ; \, 0, \, otherwise\}$.

Using Z and H we can write the one-locus model in regression form

$$Y = Za + Hd + E$$

The conditional means in this model are

- $E[Y|A_1A_1] = E[Y|Z = 1, H = 0] = a$
- $E[Y|A_1A_2] = E[Y|Z = 0, H = 1] = d$
- $E[Y|A_2A_2] = E[Y|Z = -1, H = 0] = -a$

This shows that the regression approach is equivalent than the approach we followed before (1.1). Here, we won't assume HWE and will esitmate the parameters $\{\mu, a, d\}$ variances and heritabilities using linear regression.

To fully specify the linear regression model we will relax the assumption $\mu = 0$. Using this we can now model the three genetic values using

$$Y = \mu + Za + Hd + E$$

which leads to the following conditional means

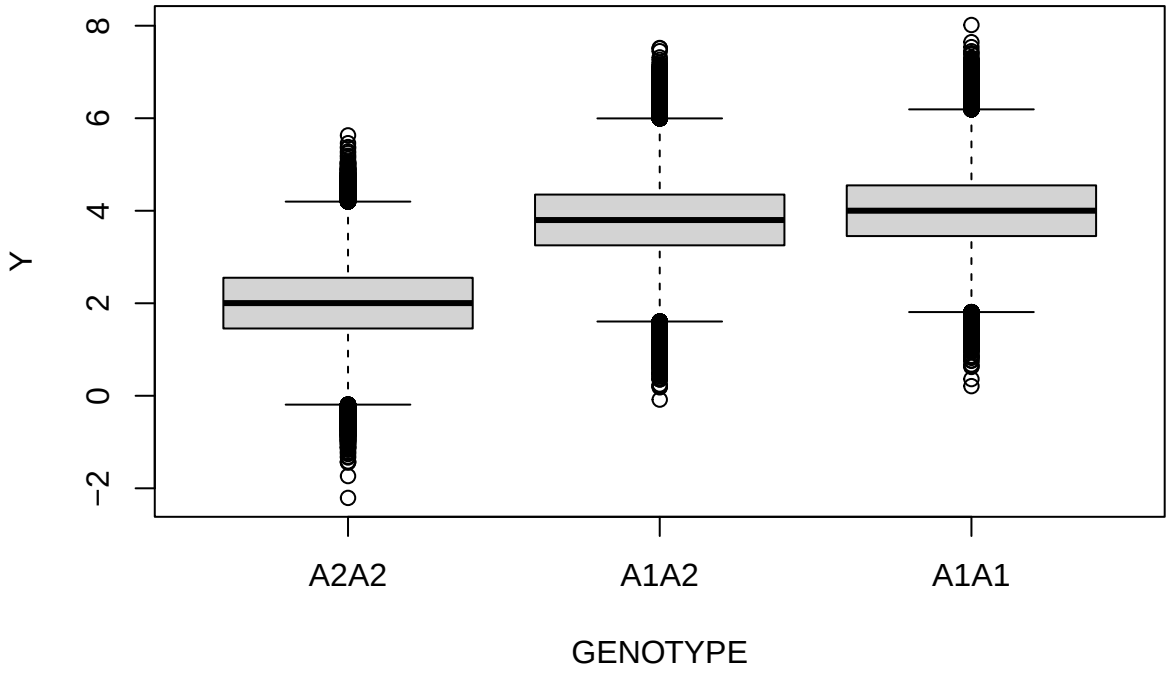
- $E[Y|A_1A_1] = E[Y|Z = 1, H = 0] = \mu + a$,
- $E[Y|A_1A_2] = E[Y|Z = 0, H = 1] = \mu + d$, and
- $E[Y|A_2A_2] = E[Y|Z = -1, H = 0] = \mu - a$.

Example 1

The following script simulates genotypes under HWE, then generate genetic values assuming $\{\mu = 3, a = 1, d = 0.8\}$ and recovers the values of the parameters through linear regression of genetic values on genotypes via ordinary least squares. For the simulation, we assume an allele frequency of 0.25 and a broad-sense heritability of 0.5. We use a large sample size to get estimates with small sampling variance.

```
# Parameters
p=0.5;  H2=0.5; n=1e6;  mu=3; a=1; d=0.8
```

```r
# Genotype and genotype covariates
X= rbinom(n=n,size=2,prob=p) # X is 0/1/2

Z=X-1
H=1*(Z==0)

# Genetic values
G=mu+Z*a+H*d

# Variances
Vg=var(G) # Genetic
Ve=Vg*((1-H2)/H2) # Environmental

# Environmental effects
E=rnorm(sd=sqrt(Ve),n=n)

# Phenotype
Y=G+E

GENOTYPE=ifelse(X==2,'A1A1',ifelse(X==1,'A1A2','A2A2'))
GENOTYPE=factor(GENOTYPE,levels=c('A2A2','A1A2','A1A1'))
boxplot(Y~GENOTYPE)
```



Estimation

```r
fm=lm(Y~Z+H)


cbind(c(mu,a,d),coef(fm))
```

```
##             [,1]      [,2]
## (Intercept)  3.0 3.0012730
## Z            1.0 0.9992904
```

```
## H           0.8 0.7991644
```

Average effect of an allele substitution (α)

The average effect of an allele (say $A1$) is defined as the expected change in genetic value that you would observe if you:

- Sample at random an individual from the population,
- Sample at random one allele from the sampled individual, and
- Replace the sampled allele with the allele in question (e.g., A_1).

For allele A_1 the allele substitution effect is denoted as $\alpha(A_1)$. Likewise, we can define $\alpha(A_2)$ to denote the average effect of substituting an allele sample at random from the population with an A_2 allele.

The **average allele substitution effect** is the difference between $\alpha(A_1)$ and $\alpha(A_2)$ and is denoted as $\alpha = \alpha_1 - \alpha_2$.

This concept plays a central role in the definition of breeding values and the additive variance.

We can also arrive at the average allele substitution effect from a linear regression perspective.

Consider regressing the genetic values on allele dosages (i.e., a covaraite that counts the number of copies of a reference allele (e.g., $A1$), that is

$G = \mu + X\alpha + \delta$

The regression coefficient is

$$\alpha = Cov(X, G)/Var(X) = Cov(X, G)/Var(X) = a + d(q - p)$$

.

We can derive α under HWE using rules of expectations, variances and (co)variances or by regressing phenotypes on allele dosage using the linear model presented above with a phenotype as the response, $Y = \mu + X\alpha + \delta$.

Regressing $Y = \mu + G + E$ on X yields an unbiased estimate of α. However, note that in this additive model, $Y = \tilde{\mu} + X\alpha + \varepsilon$, the error term ($\varepsilon$), captures environmental effects (E) plus dominance-deviations (δ), which were not captured by the linear mode. Likewise, $\tilde{\mu} = \mu + E[Zd]$.

##Example 2

```
# Large sample size and no error to minimize sampling variance
n=1e6
p=0.3; q=1-p
X=rbinom(n=n,size=2,prob=p)


a=1;d=.8
G=ifelse(X==0,-a,ifelse(X==1,d,a))

H=as.integer(X==1)

fmAD<-lm(G~X+H)
aHat=fmAD$coef[2]
dHat=fmAD$coef[3]
cbind(c(a,d),c(aHat,dHat))
```

```
##    [,1] [,2]
## X  1.0  1.0
## H  0.8  0.8
```

```
pHat=mean(X)/2
qHat=1-pHat

alpha=a+d*(q-p)
# compares true alpha, versus two estimates one derived from fmA, the other derived from fmAD
fmA<-lm(G~X)
c(alpha,fmA$coef[2],aHat+dHat*(qHat-pHat))
```

```
##                   X         X
## 1.320000 1.320400 1.320315
```

Remarks

- Regressing a phenotype on a contrast for additive effect (e.g. X={-1;0;1}) plus a contrast for dominance, H, leads yo estimates for a and d.
- Regressing a phenotype on only an additive contrast only yields an estimate of $\alpha = a + d(q - p)$.
- α can be estimated either using an additive model (`fmA` in the above example) or by estimating a and d from the regression $Y = \mu + Xa + Zd + \varepsilon$ and then plugging those into $\alpha = a + d * (q - p)$.

Additive variance

The additive variance, σ^2_α is the amount of variance that can be explained by regression on an additive contrast (X). Thus

$\sigma^2_\alpha = Var(X\alpha) = Var(X)\alpha^2 = 2pq\alpha^2$.

Here, we used the fact that under HWE, $Var(X) = E(X^2) - E(X)^2 = p^2 + q^2 - (p - q)^2 = 2pq$.

Dominance variance:

Using a sequential decomposition of the genetic variance, we can define the dominance variance as:

$\sigma^2_\gamma = \sigma^2_G - \sigma^2_\alpha = (2pqd)^2$

Single-locus Heritability

As noted earlier, using $Y = G + E$ and assuming there is no (co)variance between G and E, we can define the (single-locus) **broad-sense heritability** as the proportion of the phenotypic variance that can be explained by inter-individual differences in genetic values, that is

$H^2 = Var(G)/Var(Y) = Var(G)/[Var(G) + Var(E)] = \sigma^2_G/[\sigma^2_G + \sigma^2_E]$.

Likewise, using the 'additive model' we can define the **narrow-sense heritability** as

$h^2 = Var(X\alpha)/Var(Y) = \sigma^2_\alpha/[\sigma^2_\alpha + \sigma^2_\gamma + Var(E)]$.

Example 3

Here, we add an environmental effect with a variance such that $H^2 = 0.1$, to the simulation of Examples 1 and 2, and use linear models to estimate narrow and broad sense heritabilities for a single-locus model.

```
H2=0.1
vG=var(G)
vE=vG/H2*(1-H2)
```

```
vA=2*p*q*(alpha^2)
vG/(vG+vE)
```

```
## [1] 0.1
```

```
# 'true' narrow sense heritability
h2=vA/(vG+vE)

# note, although we have dominance, h2 is ~86% of H2
c(H2,h2,h2/H2)
```

```
## [1] 0.10000000 0.08663726 0.86637259
```

```
# Simulating a phenotype
E=rnorm(n=n,sd=sqrt(vE))
Y=100+G+E

# Fitting Additive and A+D models
fmA=lm(Y~X)
fmAD=lm(Y~X+Z)

# ANOVA
SSy=sum((Y-mean(Y))^2)
RSSa=sum(residuals(fmA)^2)
RSSad=sum(residuals(fmAD)^2)

H2Hat=(SSy-RSSad)/SSy

h2Hat=(SSy-RSSa)/SSy

cbind('Parameter'=c('H2'=H2,'h2'=h2),'Estimate'=round(c('H2'=H2Hat,'h'=h2Hat),5))
```

```
##     Parameter Estimate
## H2 0.10000000   0.0875
## h2 0.08663726   0.0875
```

Acknowledgments: I Alexa Lupi (PhD) for reading and commenting an early version of this handout.

References

Falconer, D.S. and Mackay, T.F.C. (1996) Introduction to Quantitative Genetics. 4th Edition, Addison Wesley Longman, Harlow.