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.

Thorughout the document we use a standard quantiative genetics model that views a phenotypic measurement (Y) as the sum of a genetic value (G) and an environmental effect (E)

$$Y = G + E.$$

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

In a bi-allelic locus with alleles $(A_1; A_2)$, for a diploid organism, 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 means, the variances, and heritability) using two equivalent approaches.

First, we will derive the mean and variance under Hardy-Weinberg assumptions (HW). Second, we will derive the same parameters using a regression approach. Both approaches are equivalent; however, the regression approach will pave the way towards the models commonly used in GWAS and genomic prediction.

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

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 after subtracting 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 ($d = a$), and over-dominance ($|d| > |a|$).

Technically, 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.

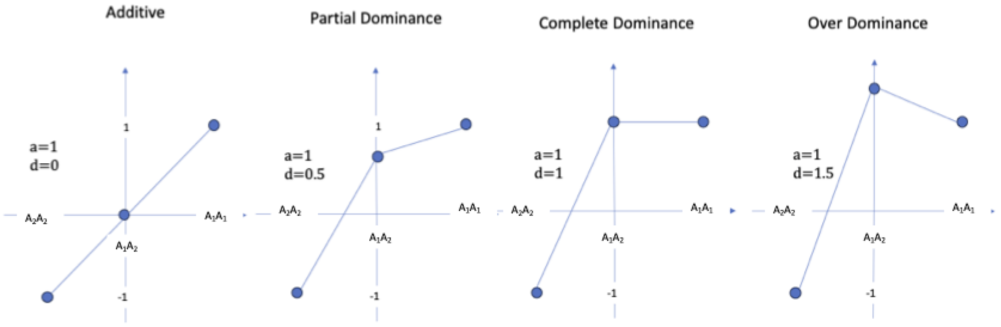

Figure 1: Figure 2: Means in the Single Locus Model

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			$a(p - q) + 2pqd$

Additive action: As noted, if $d = 0$ (strictly additive action) the genetic value of the heterozygous is exactly the average of 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:

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

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. Applying the variance function to each side of the phenotypic equation, $Y = G + E$, we get that phenotypic variance is $Var(Y) = Var(G) + Var(E) + 2Cov(G, E)$. If we assume that genetic and environmental factors are un-correlated, we have

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

Above, we partition the phenotypic variance into genetic ($Var(G)$) and environmental ($Var(E)$) components. In this setting, the proportion of the phenotypic variance explained by genetic factors (or broad sense heritability) is

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

Where, $Var(G) = 2pq(a + d[q-p])^2 + (2pqd)^2$,

Departures from HWE

Above, we derive the mean, vairance, and heritability under HWE. The same framework can be applied without making HWE assumptions by replacing the predicted probabilities of each of the genotypes with the observed ones, i.e., replacing in the tables presented above p^2, $2pq$ and q^2 with $p(A_1A_1)$, $p(A_2A_2)$, and $p(A_2A_2)$–these probabilities can be estimated from genotype data.

1.2 Regression approach

In this section we 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

defines a *contrast* between the means of A_1A_1 and A_2A_2. Additionally, we will introduce H, 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 as follows

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

The conditional means in this model are

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

Which is equivalent to the parameterization presented earlier in this document.

Example 1

Consider a single locus under HWE with the following parameters

```
# Parameters
p=0.05;   mu=3; a=1; d=0.8
```

Using the formulas derived above, $E[G] = \mu + a(p-q) + 2pqd$ and $\sigma_G^2 = 2pq(a + d[q-p])^2 + (2pqd)^2$, the mean and genetic variance at the locus are:

```
q=1-p

# Formula derived above
mG=mu+a*(p-q)+2*p*q*d
vG=2*p*q*((a+d*(q-p))^2)+(2*p*q*d)^2
message(' Mean: ',mG, ' Variance: ',vG,'.')
```

```
##  Mean: 2.176 Variance: 0.286824.
```

We now simulate a large number of genotypes under HWE, compute the genetic values using the parameters defined above and verify that the realized mean and variance are very close to the ones derived above.

```
# Genotypes and genetic values
n=1e6 # large n
X= rbinom(n=n,size=2,prob=p) # X is 0/1/2
G=ifelse(X==2,mu+a,ifelse(X==1,mu+d,mu-a))

mG.b=mean(G)
vG.b=var(G)

message('Means: mG=',mG,' ; realized mean=', round(mG.b,4),'.')
```

```
## Means: mG=2.176 ; realized mean=2.175.
```

```
message('Variance: vG=',vG,' ; realized var.=', round(vG.b,4),'.')
```

```
## Variance: vG=0.286824 ; realized var.=0.2854.
```

Recovering μ, a and d using regression analysis

```
H=ifelse(X==1,1,0)
fmAD=lm(G~X+H)
tmp=cbind(c(a,d),unname(coef(fmAD)[-1]))
rownames(tmp)=c('a','d')
colnames(tmp)=c('Moments', 'Regression')
knitr:::kable(tmp)
```

	Moments	Regression
a	1.0	1.0
d	0.8	0.8

Average effect of an allele substitution (α)

The average effect of the A_1 allele is defined as the expected change in genetic value of an individual 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 α_1. Likewise, we can define α_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 α_1 and α_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

$$\alpha = Cov(X, G)/Var(X)$$

.

We can also derive α under HWE using rules of expectations, variances and (co)variances. It can be shown that under HWE $\alpha = Cov(X, G)/Var(X) = a + d(q - p)$.

Regressing Y on X yields an unbiased estimate of α.

Example 2

```
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
## 1.720000 1.719019 1.720414
```

Remarks

- Regressing a phenotype on a contrast for additive effect (e.g. Z={-1;0;1} or X={0,1,2}) plus a contrast for dominance, H, leads yo estimates for a and d.
- Regressing a phenotype on only allele dosage yelds 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

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

Under HWE $Var(X) = 2pq$; therefore,

$$\sigma^2_\alpha = 2pq\alpha^2.$$

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 mode. Although we have a sizable dominance effect (0.8), the linear model is able to capture more than 95% of the total genetic variance (compare H^2 with h^2), this happens because the allele frequency is far from 0.5.

```r
H2=0.1
vG=var(G)
vE=vG/H2*(1-H2)
vA=2*p*q*(alpha^2)
vG/(vG+vE)
```

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

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

```
c(H2,h2,h2/H2)
```

```
## [1] 0.10000000 0.09846218 0.98462180
```

```
# 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+H)

# 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  0.09995
## h2 0.09846218  0.09792
```

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.