Means and Variances

Gustavo de los Campos (gustavoc@msu.edu)

September 8th, 2025

I. Single-locus model

This section presents the math for deriving the mean and genetic variance in a single locus model under Hardy-Weinberg equilibrium. The notation follows closely the one used Chapter 7 of Falconer & Mackay (1996).

Genotypes

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, and 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

where $q = p(A_2) = 1 - p$.

Genetic values

The genetic value of an individual (G) is the expected phenotype given the genotype $G = E[Y|Genotype]$.

Since we have three possible genotypes, for a single-locus model there are three possible 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 heterozygous, whthere this is positive or negative depends on the way we nuemrically code genotypes (i.e., which one is the reference allele). On the other hand, d is the deviation from the additive model needed to fit the genetic value of the heterozygous.

Note: To follow Falconer & Mackay notation we have coded genotypes as $\{-1, 0, 1\}$; howeve the model can equivalently be represented using allele dosages $\{0, 1, 2\}$.

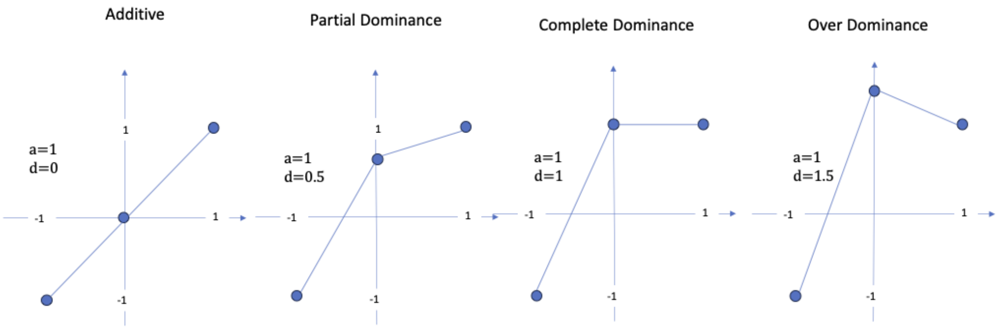


Figure 1: Figure 2: Means in the Single Locus Model

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: 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 $E[G^2]$

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:

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

Single-locus broad-sense heritability

The broad sense heritability of a trait is the proportion of the phenoptypic 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 $Var(G)$.

Regression approach

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

Consider a linear model of the form

$$Y = Za + Hd + E$$

where $Z \in \{-1,0,1\}$ and $H = \{1,\ if\ heterozygous\ ;\ 0,\ otherwise\}$.

The expected values in this model are

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

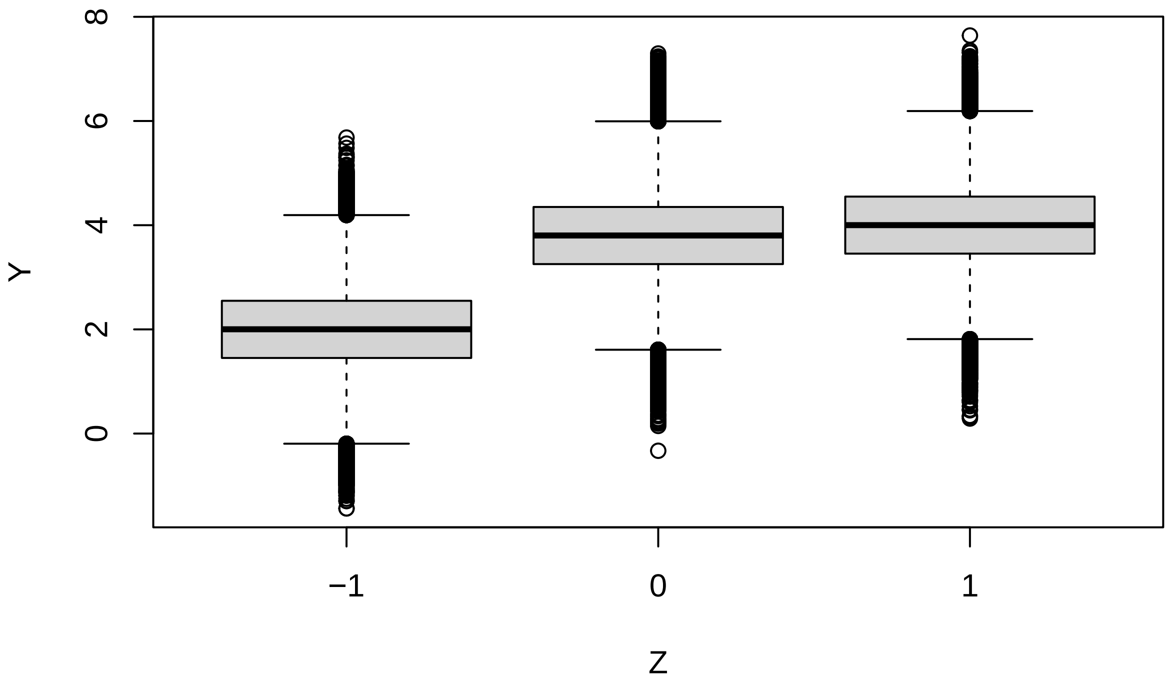
If we now relax the assumption $\mu = 0$ we can then estimate a and d through linear regression using:

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

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.

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



# 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

boxplot(Y~Z)
```



Estimation

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

summary(fm)
```

```
##
## Call:
## lm(formula = Y ~ Z + H)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1313 -0.5483  0.0005  0.5478  3.6810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.001386   0.001149  2611.8   <2e-16 ***
## Z           1.000047   0.001149   870.2   <2e-16 ***
## H           0.799723   0.001624   492.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8122 on 999997 degrees of freedom
## Multiple R-squared:  0.4997, Adjusted R-squared:  0.4997
## F-statistic: 4.994e+05 on 2 and 999997 DF,  p-value: < 2.2e-16
```

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.