Basic Designs for Estimation of Genetic Parameters

Bruce Walsh lecture notes
Liege May 2011 course
version 26 May 2011
Heritability

Narrow vs. broad sense

Narrow sense: \( h^2 = \frac{V_A}{V_P} \)

Slope of midparent-offspring regression
(sexual reproduction)

Broad sense: \( H^2 = \frac{V_G}{V_P} \)

Slope of a parent-cloned offspring regression
(asexual reproduction)

When one refers to heritability, the default is narrow-sense, \( h^2 \)

\( h^2 \) is the measure of (easily) usable genetic variation under sexual reproduction
Why $h^2$ instead of $h$?

Blame Sewall Wright, who used $h$ to denote the correlation between phenotype and breeding value. Hence, $h^2$ is the total fraction of phenotypic variance due to breeding values.

\[
r(A, P) = \frac{\sigma(A, P)}{\sigma_A \sigma_P} = \frac{\sigma_A^2}{\sigma_A \sigma_P} = \frac{\sigma_A}{\sigma_P} = h
\]

Heritabilities are functions of populations

Heritability values only make sense in the content of the population for which it was measured.

Heritability measures the *standing genetic variation* of a population. A zero heritability DOES NOT imply that the trait is not genetically determined.
Heritabilities are functions of the distribution of environmental values (i.e., the universe of E values).

Decreasing $V_p$ increases $h^2$.

Heritability values measured in one environment (or distribution of environments) may not be valid under another.

Measures of heritability for lab-reared individuals may be very different from heritability in nature.
Heritability and the prediction of breeding values

If \( P \) denotes an individual's phenotype, then best linear predictor of their breeding value \( A \) is

\[
A = \frac{\sigma(P, A)}{\sigma_P^2} (P - \mu_P) + e = h^2(P - \mu_P) + e
\]

The residual variance is also a function of \( h^2 \):

\[
\sigma_e^2 = (1 - h^2)\sigma_A^2
\]

The larger the heritability, the tighter the distribution of true breeding values around the value \( h^2(P - \mu_P) \) predicted by an individual's phenotype.
Heritability and population divergence

*Heritability is a completely unreliable predictor of long-term response*

Measuring heritability values in two populations that show a difference in their means provides no information on whether the underlying difference is genetic.
### Sample heritabilities

<table>
<thead>
<tr>
<th>Species</th>
<th>Trait</th>
<th>$h^s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>Height</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Serum IG</td>
<td>0.45</td>
</tr>
<tr>
<td>Pigs</td>
<td>Back-fat</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Weight gain</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Litter size</td>
<td>0.05</td>
</tr>
<tr>
<td>Fruit Flies</td>
<td>Abdominal Bristles</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Body size</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Ovary size</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Egg production</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Traits more closely associated with fitness tend to have lower heritabilities.
Estimation: One-way ANOVA

Simple (balanced) full-sib design: N full-sib families, each with n offspring: One-way ANOVA model

\[ Z_{ij} = \mu + f_i + w_{ij} \]

- Trait value in sib \( j \) from family \( i \)
- Common Mean
- Deviation of sib \( j \) from the family mean
- Effect for family \( i \) = deviation of mean of \( i \) from the common mean
Covariance between members of the same group equals the variance among (between) groups

\[
\text{Cov(Full Sibs)} = \sigma(z_{ij}, z_{ik}) \\
= \sigma[(\mu + f_i + w_{ij}), (\mu + f_i + w_{ik})] \\
= \sigma(f_i, f_i) + \sigma(f_i, w_{ik}) + \sigma(w_{ij}, f_i) + \sigma(w_{ij}, w_{ik}) \\
= \sigma_f^2 \\
\]

Hence, the variance among family effects equals the covariance between full sibs

\[
\sigma_f^2 = \sigma_A^2/2 + \sigma_D^2/4 + \sigma_{Ec}^2
\]
The within-family variance $\sigma_w^2 = \sigma_p^2 - \sigma_f^2$.

\[
\sigma^2_{w(FS)} = \sigma_P^2 - \left( \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_{Ec}^2 \right) \\
= \sigma_A^2 + \sigma_D^2 + \sigma_E^2 - \left( \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_{Ec}^2 \right) \\
= \left( \frac{1}{2} \right) \sigma_A^2 + \left( \frac{3}{4} \right) \sigma_D^2 + \sigma_E^2 - \sigma_{Ec}^2
\]
### One-way Anova: N families with n sibs, $T = Nn$

<table>
<thead>
<tr>
<th>Factor</th>
<th>Degrees of freedom, df</th>
<th>Sums of Squares (SS)</th>
<th>Mean sum of squares (MS)</th>
<th>$E[ MS ]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among-family</td>
<td>N-1</td>
<td>$SS_F = \sum_{i=1}^{N} n \sum_{i=1}^{n} (\bar{z}_i - \bar{z})^2$</td>
<td>$SS_f/(N-1)$</td>
<td>$\sigma^2_w + n \sigma^2_f$</td>
</tr>
<tr>
<td>Within-family</td>
<td>T-N</td>
<td>$SS_W = \sum_{i=1}^{N} \sum_{j=1}^{n} (z_{ij} - \bar{z}_i)^2$</td>
<td>$SS_w/(T-N)$</td>
<td>$\sigma^2_w$</td>
</tr>
</tbody>
</table>
Estimating the variance components:

\[ \text{Var}(f) = \frac{\text{MS}_f - \text{MS}_w}{n} \]

\[ \text{Var}(w) = \text{MS}_w \]

\[ \text{Var}(z) = \text{Var}(f) + \text{Var}(w) \]

Since \( \sigma_f^2 = \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_{Ec}^2 \)

\(2\text{Var}(f)\) is an upper bound for the additive variance
Assigning standard errors ( = square root of Var)

Fun fact: Under normality, the (large-sample) variance for a mean-square is given by

\[
\sigma^2(MS_x) \approx \frac{2(\text{MS}_x)^2}{df_x + 2}
\]

\[
\text{Var}[\text{Var}(w(\text{FS}))] = \text{Var}(\text{MS}_w) \approx \frac{2(\text{MS}_w)^2}{T - N + 2}
\]

\[
\text{Var}[\text{Var}(f)] = \text{Var} \left[ \frac{\text{MS}_f - \text{MS}_w}{n} \right] \\
\approx \frac{2}{n^2} \left( \frac{(\text{MS}_f)^2}{N + 1} + \frac{(\text{MS}_w)^2}{T - N + 2} \right)
\]
Estimating heritability

Hence, $h^2 < 2 \, t_{FS}$

An approximate large-sample standard error for $h^2$ is given by

$$SE(h^2) \approx 2(1 - t_{FS})[1 + (n - 1)t_{FS}] \sqrt{2/[Nn(n-1)]}$$
Worked example

10 full-sib families, each with 5 offspring are measured

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among-families</td>
<td>9</td>
<td>SS_{f} = 405</td>
<td>45</td>
<td>\sigma^2_{w} + 5 \sigma^2_{f}</td>
</tr>
<tr>
<td>Within-families</td>
<td>40</td>
<td>SS_{w} = 800</td>
<td>20</td>
<td>\sigma^2_{w}</td>
</tr>
</tbody>
</table>

\[
\text{Var}(f) = \frac{\text{MS}_f - \text{MS}_w}{n} = \frac{45 - 20}{5} = 5 \quad \longrightarrow \quad V_A < 10
\]

\[
\text{Var}(w) = \text{MS}_w = 20
\]

\[
\text{Var}(z) = \text{Var}(f) + \text{Var}(w) = 25
\]

\[
\text{SE}(h^2) \simeq 2(1 - 0.4)[1 + (5 - 1)0.4]\sqrt{2/[50(5 - 1)]} = 0.312
\]
Full sib-half sib design: Nested ANOVA

Full-sibs

Half-sibs
Estimation: Nested ANOVA

Balanced full-sib / half-sib design: N males (sires) are crossed to M dams each of which has n offspring:

Nested ANOVA model

\[ z_{ijk} = \mu + s_i + d_{ij} + w_{ijk} \]

Value of the kth offspring from the jth dam for sire i

Overall mean

Effect of sire i = deviation of mean of i's family from overall mean

Effect of dam j of sire i = deviation of mean of dam j from sire and overall mean

Within-family deviation of kth offspring from the mean of the ij-th family
Nested ANOVA model:

\[ Z_{ijk} = \mu + S_i + d_{ij} + w_{ijk} \]

\[ \sigma^2_s = \text{between-sire variance} = \text{variance in sire family means} \]

\[ \sigma^2_d = \text{variance among dams within sires} = \text{variance of dam means for the same sire} \]

\[ \sigma^2_w = \text{within-family variance} \]

\[ \sigma^2_T = \sigma^2_s + \sigma^2_d + \sigma^2_w \]
Nested Anova: N sires crossed to M dams, each with n sibs, T = NMn

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sires</td>
<td>N-1</td>
<td>[ SS_s = Mn \sum_{i=1}^{N} \sum_{j=1}^{M_i} (\bar{z}_i - \bar{z})^2 ]</td>
<td>[ SS_d/(N-1) ]</td>
<td>[ \sigma_w^2 + n\sigma_d^2 + Mn\sigma_s^2 ]</td>
</tr>
<tr>
<td>Dams(Sires)</td>
<td>N(M-1)</td>
<td>[ SS_d = n \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{z}_{ij} - \bar{z}_i)^2 ]</td>
<td>[ SS_d/(N[M-1]) ]</td>
<td>[ \sigma_w^2 + n\sigma_d^2 ]</td>
</tr>
<tr>
<td>Sibs(Dams)</td>
<td>T-NM</td>
<td>[ SS_w = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n} (\bar{z}<em>{ijk} - \bar{z}</em>{ij})^2 ]</td>
<td>[ SS_w/(T-NM) ]</td>
<td>[ \sigma_w^2 ]</td>
</tr>
</tbody>
</table>
Estimation of sire, dam, and family variances:

\[
\text{Var}(s) = \frac{MS_s - MS_d}{Mn}
\]

\[
\text{Var}(d) = \frac{MS_d - MS_w}{n}
\]

\[
\text{Var}(e) = MS_w
\]

Translating these into the desired variance components

- \(\text{Var(Total)} = \text{Var(between FS families)} + \text{Var(Within FS)}\)

  \[\rightarrow \sigma^2_w = \sigma^2_z - \text{Cov(FS)}\]

- \(\text{Var(Sires)} = \text{Cov(Paternal half-sibs)}\)

  \[\sigma^2_d = \sigma^2_z - \sigma^2_s - \sigma^2_w = \sigma(FS) - \sigma(\text{PHS})\]
Summarizing,

\[
\sigma^2_s = \sigma(PHS) \quad \sigma^2_d = \sigma^2_z - \sigma^2_s - \sigma^2_w = \sigma(FS) - \sigma(PHS)
\]

\[
\sigma^2_w = \sigma^2_z - \sigma(FS) \quad \sigma^2_s = \sigma^2_z - \sigma^2_s - \sigma^2_w = \sigma(FS) - \sigma(PHS)
\]

Expressing these in terms of the genetic and environmental variances,

\[
\sigma^2_s \approx \frac{\sigma^2_A}{4}
\]

\[
\sigma^2_d \approx \frac{\sigma^2_A}{4} + \frac{\sigma^2_D}{4} + \sigma^2_{Ec}
\]

\[
\sigma^2_w \approx \frac{\sigma^2_A}{2} + \frac{3\sigma^2_D}{4} + \sigma^2_{Es}
\]
Intraclass correlations and estimating heritability

\[ t_{PHS} = \frac{\text{Cov}(PHS)}{\text{Var}(z)} = \frac{\text{Var}(s)}{\text{Var}(z)} \quad \rightarrow \quad 4t_{PHS} = h^2 \]

\[ t_{FS} = \frac{\text{Cov}(FS)}{\text{Var}(z)} = \frac{\text{Var}(s) + \text{Var}(d)}{\text{Var}(z)} \quad \rightarrow \quad h^2 \leq 2t_{FS} \]

Note that \( 4t_{PHS} = 2t_{FS} \) implies no dominance or shared family environmental effects.
Worked Example:
N=10 sires, M = 3 dams, n = 10 sibs/dam

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sires</td>
<td>9</td>
<td>4,230</td>
<td>470</td>
<td>$\sigma_w^2 + 10\sigma_d^2 + 30\sigma_s^2$</td>
</tr>
<tr>
<td>Dams(Sires)</td>
<td>20</td>
<td>3,400</td>
<td>170</td>
<td>$\sigma_w^2 + 10\sigma_d^2$</td>
</tr>
<tr>
<td>Within Dams</td>
<td>270</td>
<td>5,400</td>
<td>20</td>
<td>$\sigma_w^2$</td>
</tr>
</tbody>
</table>

\[
\sigma_w^2 = MS_w = 20
\]
\[
\sigma_d^2 = \frac{MS_d - MS_w}{n} = \frac{170 - 20}{10} = 15
\]
\[
\sigma_s^2 = \frac{MS_s - MS_d}{NN} = \frac{470 - 170}{30} = 10
\]
\[
\sigma_P^2 = \sigma_s^2 + \sigma_d^2 + \sigma_w^2 = 45
\]
\[
\sigma_d^2 = 15 = (1/4)\sigma_A^2 + (1/4)\sigma_D^2 + \sigma_{Ec}^2
\]
\[
= 10 + (1/4)\sigma_D^2 + \sigma_{Ec}^2
\]
\[
\sigma_A^2 = 4\sigma_s^2 = 40
\]
\[
\sigma_D^2 + 4\sigma_{Ec}^2 = 20
\]
\[
h^2 = \frac{\sigma_A^2}{\sigma_P^2} = \frac{40}{45} = 0.89
\]
Parent-offspring regression

Single parent - offspring regression

\[ z_{oi} = \mu + b_{o|p}(z_{pi} - \mu) + e_i \]

The expected slope of this regression is:

\[
E(b_{o|p}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \approx \frac{\sigma^2_A/2 + \sigma(E_o, E_p)}{\sigma^2_z} = \frac{h^2}{2} + \frac{\sigma(E_o, E_p)}{\sigma^2_z}
\]

Residual error variance (spread around expected values)

\[
\sigma^2_e = \left(1 - \frac{h^2}{2}\right) \sigma^2_z
\]
The expected slope of this regression is:

\[ E(b_o|p) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \approx \frac{(\sigma^2_A/2) + \sigma(E_o, E_p)}{\sigma^2_z} = \frac{h^2}{2} + \frac{\sigma(E_o, E_p)}{\sigma^2_z} \]

Shared environmental values

To avoid this term, typically regressions are male-offspring, as female-offspring more likely to share environmental values.
Midparent - offspring regression

\[ z_{oi} = \mu + b_{o\|MP} \left( \frac{z_{mi} + z_{fi}}{2} - \mu \right) + e_i \]

\[ b_{o\|MP} = \frac{\text{Cov}[z_o, (z_m + z_f)/2]}{\text{Var}[(z_m + z_f)/2]} \]

\[ = \frac{[\text{Cov}(z_o, z_m) + \text{Cov}(z_o, z_f)]/2}{[\text{Var}(z) + \text{Var}(z)]/4} \]

\[ = \frac{2\text{Cov}(z_o, z_p)}{\text{Var}(z)} = 2b_{o|p} \]

The expected slope of this regression is \( h^2 \)

Residual error variance (spread around expected values)

\[ \sigma_e^2 = \left( 1 - \frac{h^2}{2} \right) \sigma_z^2 \]
Standard errors

Single parent-offspring regression, $N$ parents, each with $n$ offspring

Squared regression slope

$$\text{Var}(b_o|p) \approx \frac{n(t - b_o^2|p) + (1 - t)}{Nn}$$

Total number of offspring

Sib correlation

$$t = \begin{cases} 
    t_{HS} = h^2/4 & \text{for half-sibs} \\
    t_{FS} = h^2/2 + \frac{\sigma_D^2 + \sigma_E^2}{\sigma_z^2} & \text{for full sibs}
\end{cases}$$

$$\text{Var}(h^2) = \text{Var}(2b_o|p) = 4\text{Var}(b_o|p)$$
Midparent-offspring regression, 
N sets of parents, each with n offspring

\[
\text{Var}(h^2) = \text{Var}(b_o|_{MP}) \approx \frac{2[n(t_{FS} - b_o|_{MP}/2) + (1 - t_{FS})]}{Nn}
\]

- Midparent-offspring variance half that of single parent-offspring variance

\[
\text{Var}(h^2) = \text{Var}(2b_o|_p) = 4\text{Var}(b_o|_p)
\]
Estimating Heritability in Natural Populations

Often, sibs are reared in a laboratory environment, making parent-offspring regressions and sib ANOVA problematic for estimating heritability.

Let $b'$ be the slope of the regression of the values of lab-raised offspring regressed in the trait values of their parents in the wild.

A lower bound can be placed of heritability using parents from nature and their lab-reared offspring, 

$$h_{min}^2 = (b'_o|MP)^2 \frac{\text{Var}_n(z)}{\text{Var}_l(A)}$$

- Trait variance in nature
- Additive variance in lab
Why is this a lower bound?

\[
(b'_{o|MP})^2 \frac{\text{Var}_n(z)}{\text{Var}_l(A)} = \left[ \frac{\text{Cov}_{l,n}(A)}{\text{Var}_n(z)} \right]^2 \frac{\text{Var}_n(z)}{\text{Var}_l(A)} = \gamma^2 h^2_n
\]

where

\[
\gamma = \frac{\text{Cov}_{l,n}(A)}{\sqrt{\text{Var}_n(A)\text{Var}_l(A)}}
\]

is the additive genetic covariance between environments and hence \(\gamma^2 \leq 1\).
Defining $H^2$ for Plant Populations

Plant breeders often do not measure individual plants (especially with pure lines), but instead measure a plot or a block of individuals. This can result in inconsistent measures of $H^2$ even for otherwise identical populations.

$$z_{ijk\ell} = G_i + E_j + GE_{ij} + p_{ijk} + e_{ijk\ell}$$

- $z_{ijk\ell}$: deviations of individual plants within this plot
- $G_i$: Effect of plot $k$ for Genotype $i$ in environment $j$
- $E_j$: Genotype $i$ and environment $j$
- $GE_{ij}$: Interaction between Genotype $i$ and environment $j$
- $p_{ijk}$: Environment $j$
- $e_{ijk\ell}$: Genotype $i$
Hence, $V_p$, and hence $H^2$, depends on our choice of $e$, $r$, and $n$.

\[ z_{ijkl} = G_i + E_j + GE_{ij} + p_{ijk} + e_{ijkl} \]

\[ \sigma^2(z_i) = \sigma^2_G + \sigma^2_E + \frac{\sigma^2_{GE}}{e} + \frac{\sigma^2_p}{er} + \frac{\sigma^2_e}{ern} \]

- $e = \text{number of environments}$
- $r = \text{(replicates) number of plots/environment}$
- $n = \text{number of individuals per plot}$