Estimating population sizes for breeding and genetics, Continued.

As we continue our discussion of estimating family sizes for breeding and genetics applications, I would like to review the last material covered in lecture 11.

What family size is required to obtain estimates of genetic variances?

This question is relevant to experiments designed to measure genetic variability (or heritability). The experimental design seeks to measure phenotypic variability \( (M_1 = V_p) \) with \( f_1 \) degrees of freedom. Phenotypic variability contains both \( V_g \) and environmental variability \( (M_2 = V_e) \) with \( f_2 \) degrees of freedom. For a simple experiment \( V_g \) is a linear combination of two mean squares. In an experiment where \( f_1 + 1 \) genotypes are measured “r” times

\[
V_g = (V_p - V_e)/r
\]

(This equation should be familiar in the context of the Cotterril paper that we read earlier in class).

Note that heritability = \( V_g/(V_e + V_g) \)

The standard error of \( V_g \) can be expressed as a proportion of \( V_g \) using the relationship

\[
[(V_g(V_g)/V_g)^{1/2} = [2((b(V_p/V_e)^2) + 1)/bf1((V_p/V_e) – 1)^2]^{1/2}
\]

where \( bf1 = f_2 \) so \( b = f_2/f_1 \) and \( b \) is a constant dependent on the degrees of freedom for phenotypic variability and environmental variability.

References:


Number of replicates that minimize the variance for \( V_g \) at given levels of heritability.

<table>
<thead>
<tr>
<th>No. of Plots</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>low</td>
<td>moderate</td>
<td>high</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>300</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Take home message: More replication is required for low heritability traits and large numbers of genotypes are required to minimize variation.
Genotypes (df for phenotype mean square) required to yield proportionate standard errors for Vg at selected levels of heritability

\[
\frac{\text{Vg}}{\text{Vg} + \text{Ve}} \quad \left( \frac{\text{V(Vg)}}{\text{Vg}} \right)^{1/2}
\]

<table>
<thead>
<tr>
<th>reps</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘1/2’</td>
<td>2</td>
<td>182</td>
<td>49</td>
<td>29</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>82</td>
<td>51</td>
<td>28</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>‘1/4’</td>
<td>2</td>
<td>728</td>
<td>416</td>
<td>194</td>
<td>116</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>325</td>
<td>203</td>
<td>111</td>
<td>76</td>
<td>59</td>
</tr>
<tr>
<td>‘1/8’</td>
<td>2</td>
<td>2910</td>
<td>1670</td>
<td>776</td>
<td>464</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1300</td>
<td>811</td>
<td>444</td>
<td>304</td>
<td>235</td>
</tr>
</tbody>
</table>

Take home messages: for most estimates of heritability from line means, the standard error will be high (25%-50% of Vg for high heritability traits). Also, if the goal is to reduce error, replication is more “efficient” for low heritability traits while adding genotypes is more “efficient” for high heritability traits. To illustrate this point, note the paired comparisons in the table below.

<table>
<thead>
<tr>
<th>Heritability</th>
<th>ST Error</th>
<th>genotypes</th>
<th>reps</th>
<th>Total plots</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>50%</td>
<td>49</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>28</td>
<td>3</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>194</td>
<td>2</td>
<td>388</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>111</td>
<td>3</td>
<td>333</td>
</tr>
<tr>
<td>70%</td>
<td>50%</td>
<td>13</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>11</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>49</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>43</td>
<td>3</td>
<td>129</td>
</tr>
</tbody>
</table>
Genotypes (df for pheonotype mean square) required to yield proportionate standard errors for Vg at selected levels of heritability (2 reps, 2 locations, 2 years)

\[
\begin{array}{c|ccccc}
\text{[(V(Vg)/Vg)}^{1/2} & 10 & 15 & 20 & 30 & 50 \\
\hline
1/2 & 67 & 36 & 24 & 16 & 11 \\
1/4 & 266 & 142 & 96 & 62 & 42 \\
1/8 & 1070 & 567 & 384 & 247 & 168 \\
\end{array}
\]

Take home message: replication over years and locations will reduce the number of plots per year, but will require more time to obtain the same accuracy of estimate. Recall back to our discussion of heritability from line-means and the importance of GxL and GxY interactions and this approach may offer an appropriate compromise.

<table>
<thead>
<tr>
<th>Heritability</th>
<th>ST Error</th>
<th>genotypes</th>
<th>reps</th>
<th>Total plots</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>50%</td>
<td>49</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>50%</td>
<td>28</td>
<td>3</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>16</td>
<td>8</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>194</td>
<td>2</td>
<td>388</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>111</td>
<td>3</td>
<td>333</td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>62</td>
<td>8</td>
<td>496</td>
<td></td>
</tr>
</tbody>
</table>

Estimating heritability with an error of 25% of Vg can be accomplished over two years and two locations with 32% of the genotypes, but 128% of the resources needed for a one-year two-replicate trial.

One final point relative to genotypes needed to minimize the error for genotypic variance (Vg or \(\sigma^2(G)\)):

The standard error for H is (from Lecture 5):

\[
SE(H) = \frac{SE\sigma^2(G)}{\sigma^2(G) + \sigma^2ge/e + \sigma^2error/r*e}
\]


How many samples must be measured to distinguish the best varieties or lines from the worst?

For the methodology used to answer this final question, we will discuss an approach outlined in a paper by E. J. Sacks and D.V. Shaw (1994, Optimum allocation of objective color measurements for evaluating fresh strawberries. J. Amer. Soc. Hort. Sci. 119:330-334).
When planning experiments designed to measure quantitative variation the concept of reproducibility for an estimation becomes useful. The reproducibility of a result is described by the Confidence Interval (CI), indicating the limits that enclose a mean.

\[
CI = (\text{critical value}) \times \frac{S}{(N)^{\frac{1}{2}}}
\]

For example for a large population 95% CI = 1.96 * S / (N)^{\frac{1}{2}} where \( t \) is the from the t-distribution at \( P = 0.95 \) for the one tailed test (0.975 for the two tailed test) at \( df = \infty \).


Future individual value \( CI = a + bX_0 \pm t(\text{MSE}[1 + 1/n + (X_0 - X)^2 / SS(X)])^{\frac{1}{2}} \)

Future mean value \( CI = a + bX_0 \pm t(\text{MSE}[1/n + (X_0 - X)^2 / SS(X)])^{\frac{1}{2}} \)


\[
N = t^2 \times \frac{S^2}{d^2}
\]

The value \( t \) is the tabulated \( t \) value for the desired confidence level and the degrees of freedom of the initial sample, \( d \) is the half-width of the desired confidence interval (i.e. \( d = CI / 2 \)), and \( S \) is the standard deviation (\( S^2 \) is the variance). Use of \( t = 1.96 \) assumes \( df \) larger than 120, in practice these will usually need to be adjusted down especially for GY.

Consider a simple experiment consisting of treatments (or genotypes), blocks per year, and years. The relationship \( N = t^2 \times \frac{S^2}{d^2} \) can be re-written as:

\[
d^2 = 1.96^2 (\sigma^2 \text{ error})/BY + 1.96^2(\sigma^2 \text{ GBY})/BY + 1.96^2 (\sigma^2 \text{ GY})/Y
\]

where \( B = \text{Blocks} \quad Y = \text{Years} \quad G = \text{Treatment (genotype)} \)

And the total treatment variance = \( \sigma^2 \text{ GY} + \sigma^2 \text{ GBY} + \sigma^2 \text{ error} \)

The equation \( d^2 = 1.96^2 (\sigma^2 \text{ error})/BY + 1.96^2(\sigma^2 \text{ GBY})/BY + 1.96^2 (\sigma^2 \text{ GY})/Y \) can be re-arranged to solve for \( B \) or \( Y \) at a given value of \( d \) (\( CI/2 \)). In other words, this relationship can be used to determine the number of \( B \) or \( Y \) needed to detect a given value of \( d \) at various probabilities (determined by the \( t \) value). The variance components can be calculated from the ANOVA (using the SAS proc varcomp or SAS proc mixed procedures).

**EXAMPLE**

Sacks and Shaw used this approach to ask the question “…how many fruit (n) [are] necessary to estimate a genotypes mean color trait value for a single [harvest] date within d units (CIELAB or degrees) with 95% confidence…”
Consider the expected mean squares for the analysis of color in their design (from Table 1):

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>External</th>
<th>Internal</th>
<th>Expected mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
<td>46</td>
<td>46</td>
<td></td>
<td>$\sigma^2 e + k1\sigma^2 f(gd) + k2\sigma^2 gd + k3\sigma^2 g$</td>
</tr>
<tr>
<td>Date</td>
<td>2</td>
<td>2</td>
<td></td>
<td>$\sigma^2 e + k1\sigma^2 f(gd) + k4\sigma^2 gd + k5\sigma^2 d$</td>
</tr>
<tr>
<td>G X D</td>
<td>88</td>
<td>87</td>
<td></td>
<td>$\sigma^2 e + k1\sigma^2 f(gd) + k6\sigma^2 gd$</td>
</tr>
<tr>
<td>Fruit within G&amp; D</td>
<td>1886</td>
<td>1871</td>
<td></td>
<td>$\sigma^2 e + k1\sigma^2 f(gd)$</td>
</tr>
<tr>
<td>Error (w/in Fruit)</td>
<td>2022</td>
<td>2006</td>
<td></td>
<td>$\sigma^2 e$</td>
</tr>
</tbody>
</table>

The coefficients k1 – k6 are estimated due to the fact that the data set is not completely balanced.

The equation used to determine N is:  

$$N = \frac{1.96^2 [\sigma^2 f(gd) + (\sigma^2 e)/R]}{d^2}$$

Note that the equation is restricted to a single harvest date so that the fruit within genotype and date variance is used and the error variance is corrected for the number of measurements per fruit, R. Also note that the df are large for error and f(gd) in the experiment, so using the t value for df = infinity is acceptable (most tables list t values for df up to ~ 120).

Questions:
- How would this equation differ if we wanted to estimate the genotypes mean with 99% confidence or 70% confidence?
- How would the equation differ if we wanted to estimate the genotypes mean over both harvest dates?

Example 2:
To help you understand how the Confidence Interval approach can be used to help plan experiments. We can consider the gene expression data set (chsdata2.xls) from Lecture 1. Remember this data set contains four treatments, three reps per treatment, and two experiments. We considered rep and experiment to be random. How can we use the variance components to determine how many reps or experiments are necessary to detect a treatment difference of 0.4 ?
Note that if you are having trouble calculating the expected mean square, you can use SAS as a shortcut. We used the following sas code to generate the MS:

Proc glm;
   Class trt exp rep;
   Model chs = trt|exp|rep(exp);
   Random trt|exp|rep(exp) / test;

Second, the general equation from above
d^2 = 1.96^2(\sigma^2 \text{ error})/BY + 1.96^2(\sigma^2 \text{ GBY})/BY + 1.96^2(\sigma^2 \text{ GY})/Y

can be re-written for our data set as follows:
d^2 = 2.179^2(\sigma^2 \text{ error})/RepExp + 3.182^2(\sigma^2 \text{ trt*exp})/Exp

Where the value of t is adjusted to the degrees of freedom for each of the effects in our model (3 df for trt*exp, and 12 df for error).

The variance components can be calculated using the proc varcmp or proc mixed procedure in SAS.

data chsdat;
    infile 'a:chsdata2.csv' delimiter = ',' firstobs = 2;
    input trt exp time rep chs;
proc sort;
    by trt exp rep;
proc mixed data = chsdat covtest;
    class trt exp rep;
    model chs = trt / ddfm=satterth;
    random exp rep(exp) trt*exp;
    title 'variance components for gene expression study trt = fixed';
run;
The Mixed Procedure

Model Information

Data Set WORK.CHSDAT
Dependent Variable chs
Covariance Structure Variance Components
Estimation Method REML
Residual Variance Method Profile
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Satterthwaite

Class Level Information

Class Levels Values
trt 4 11 12 21 22
exp 2 1 2
rep 3 1 2 3

Dimensions

Covariance Parameters 4
Columns in X 5
Columns in Z 16
Subjects 1
Max Obs Per Subject 24
Observations Used 24
Observations Not Used 0
Total Observations 24

Iteration History

Iteration Evaluations -2 Res Log Like Criterion
0 1 10.96928328
1 3 9.78181389 0.00000990
2 1 9.78167896 0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Standard Z
Cov Parm Estimate Error Value Pr Z
exp 0 . . .
rep(exp) 0.000899 0.01160 0.08 0.4691
trt*exp 0.02332 0.03061 0.76 0.2231
Residual 0.05595 0.02291 2.44 0.0073

Fitting Information

Res Log Likelihood -4.9
Akaike's Information Criterion -7.9
Schwarz's Bayesian Criterion -5.9
-2 Res Log Likelihood 9.8

Type 3 Tests of Fixed Effects

Effect DF DF F Value Pr > F
trt 3 3.93 1.61 0.3216
Asside: Using Proc Varcomp method = REML, and Proc Mixed will return approximately the same estimates for the variance components because the data set is balanced. Variance components can only be estimated for random effects.

<table>
<thead>
<tr>
<th></th>
<th>Proc Varcomp</th>
<th>Proc Mixed Trt = random</th>
<th>Proc Mixed Trt = fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt</td>
<td>0.01287</td>
<td>0.01287</td>
<td>not calculated</td>
</tr>
<tr>
<td>Exp</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rep(Exp)</td>
<td>0.0008989</td>
<td>0.000899</td>
<td>0.000899</td>
</tr>
<tr>
<td>Trt*Exp</td>
<td>0.02333</td>
<td>0.02332</td>
<td>0.02332</td>
</tr>
<tr>
<td>Error</td>
<td>0.05595</td>
<td>0.05595</td>
<td>0.05595</td>
</tr>
</tbody>
</table>

Now we have all the pieces we need to solve the problem. A useful approach to the is to write a MACRO in Excel to calculate the effect of adding Reps and Experiments on the value of d.

**EXCEL MACRO**

\[
= \sqrt{(2.179 \times 2.179 \times 0.05595)/(R \times E)) + ((3.182 \times 3.182 \times 0.02332)/(E))}
\]

R and E are replaced by the corresponding column and row identification and the MACRO is copied and pasted down the column to yield:

<table>
<thead>
<tr>
<th>R</th>
<th>E</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.708358</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.500885</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.408971</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.354179</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.607408</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.429502</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.350687</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.303704</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.569797</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.402907</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.328973</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.284899</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.550028</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.388929</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.317559</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.275014</td>
</tr>
</tbody>
</table>
The table above can be rearranged as:

<table>
<thead>
<tr>
<th>EXP</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>REP</td>
<td>1</td>
<td>0.7084</td>
<td>0.5009</td>
<td>0.4090</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6074</td>
<td>0.4295</td>
<td>0.3507</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.5698</td>
<td>0.4029</td>
<td>0.3290</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.5500</td>
<td>0.3889</td>
<td>0.3176</td>
</tr>
</tbody>
</table>

Note that our CHS gene expression data had a maximum treatment difference of 1.4235 – 1.0508 = 0.3727. So, adding a third experiment might allow us to distinguish these treatments at p < 0.05 (assuming adding a third experiment does not inflate the error variance or treatment*experiment variance). Also note that this approach is approximate given that adding reps or experiments will change the df and thus the t value when experiments are preformed with few treatments.

I have given you a gene expression example to emphasize the general applicability of this approach. However, you could apply this approach to the lycopene data set to help refine field sampling for product quality.

Finally, consider the relationship between “heritability” equations from the Cotterill paper and the generalized equation from the Confidence Interval.

\[ d^2 = t^2 \left( \frac{\sigma^2}{\text{error}} \right) /BY + t^2 \left( \frac{\sigma^2}{\text{GBY}} \right) /BY + t^2 \left( \frac{\sigma^2}{\text{GY}} \right) /Y \]

\[ H_G = \frac{\sigma^2 G}{\sigma^2 G + (\sigma^2 /\text{error}) /BY + (\sigma^2 /\text{GBY}) /BY + (\sigma^2 /\text{GY}) /Y} \]

The implications should be clear, the more environmental variance associated with a trait, the more replication is needed to distinguish varieties or treatments that differ by some value \( d = \text{CI}/2 \).
E. J. Sacks and D.V. Shaw also use the concept of “repeatability” to make decisions about how to efficiently measure a trait. The methodology we discussed above is an alternative to repeatability. Note the relationship between repeatability and heritability.

**Uses of Variance Partitioning (From Falconer):**

<table>
<thead>
<tr>
<th>Data</th>
<th>Partition</th>
<th>Ratio Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resemblance between relatives</td>
<td>$Va:Vna + Veg + Ves$</td>
<td>heritability $Va/Vp$</td>
</tr>
<tr>
<td>Genetically uniform group</td>
<td>$Va+Vna:Veg+Ves$</td>
<td>degree of genetic determination</td>
</tr>
<tr>
<td>Multiple measures</td>
<td>$Vg+Veg:Ves$</td>
<td>repeatability</td>
</tr>
</tbody>
</table>

$a =$ additive  
eg = general environment (not easily measured or separated from $Vg$)  
$na =$ non-additive  
es = special environment

**References:**


Assignment:

Read for Wednesday Discussion:


Read for familiarity:


Begin reviewing chapters in Lynch and Walsh that relate to genetic mapping. We will cover topics from the following chapters over the next ~three weeks.

9. Analysis of Line Crosses (quantitative behavior in crosses derived from inbred lines)

12. Polygenes and Polygenic Mutation (esp. 321-328; the nature of quantitative traits)

13. Detecting Major Genes

14. Principles of Marker-Based Analysis

15. Mapping and Characterizing QTLS: Inbred Line Crosses

16. Mapping and Characterizing QTLS: Outbred Populations