Understanding Heritability

Andrey Vega Alfaro

2024-02-16

Heritability measures the extent to which differences in observed phenotypes can be attributed to genetic variation and ranges between 0 and 1. Heritability has several definitions. A general definition of heritability is *the squared correlation between genotype and phenotype*. That definition is usually expressed as $H^2 = [corr(g, y)]^2$.

Plot-based Heritability

Let's see how we can get to the familiar formula for plot-based heritability that involves the ratio of genetic variance to phenotypic variance $H^2 = \frac{\sigma_g^2}{\sigma_y^2}$. Our starting point will be the definition stating that heritability represents the squared correlation between genotype and phenotype. We will assume a completely randomized design with balanced data.

Correlation definition The *correlation* between the genotype (g) and phenotype (y) is defined as the covariance between g and y divided by the product of their standard deviations:

$$corr(g, y) = \frac{cov(g, y)}{\sigma_g \sigma_y}$$

Phenotype representation Remember that phenotype (y) is represented as the sum of genotype (g) and a non genetic-signal (r), and expressed as $y = g + r$. The non-genetic signal part r is the residual error. By substituting the value of y, the correlation formula becomes:

$$corr(g, y) = \frac{cov(g, g + r)}{\sigma_g \sigma_y}$$

The additive law of covariance The additive law of covariance states that covariance of a random variable with a sum of random variables is just the sum of the covariances with each of the random variables. We have a term that is the covariance of g with a sum of $g + r$, expressed in the numerator as $cov(g, g + r)$. Using the additive law of covariance we have that the covariance of g with the sum of $g + r$ is the sum of the covariances. Our formula becomes:

$$corr(g, y) = \frac{cov(g, g) + cov(g, r)}{\sigma_g \sigma_y}$$

Variance identification We know that the covariance of a given variable with itself is called **variance**. So our formula becomes:

$$corr(g, y) = \frac{var(g, g) + cov(g, r)}{\sigma_g \sigma_y}$$

As is usual with the residual error term, we assume $r \sim N(0, \sigma_r)$. In simple terms this means that the residual errors are normally distributed and come at random with mean zero and variance σ_r. In a field setting this assumption means that the phenotypic value is estimated for all the genotypes with the same precision. We don't expect some genotypes to have more error than others. This is possible because genotypes are randomized. In consequence, we can assume that the residual errors (r) are independent of the genotype (g). This means there is no covariance between them: $cov(g, r) = 0$. The formula simplifies to:

$$corr(g, y) = \frac{var(g, g)}{\sigma_g \sigma_y}$$

Change in notation and simplify We can express $var(g, g)$ equivalently as σ_g^2 and our formula is:

$$corr(g, y) = \frac{\sigma_g^2}{\sigma_g \sigma_y}$$

This simplifies to a familiar expression.

$$corr(g, y) = \frac{\sigma_g}{\sigma_y}$$

Squaring the Correlation Broad-sense heritability on a plot basis is equal to the squared correlation. We square the formula and it becomes:

$$H^2_{plot} = [cor(g, y)]^2 = \frac{\sigma_{g^2}}{\sigma_y^2}$$

So, broad-sense heritability on a plot basis is equal to the squared correlation between phenotype and genotype and is expressed as:

$$H^2_{plot} = \frac{\sigma_{g^2}}{\sigma_y^2}$$

Variance partitioning Note that the variance of a random variable can be *partitioned* into the variances among the two components plus two times the covariance between the components. So, if $y = g + r$ then the variance can be partitioned as:

$$\sigma_y^2 = \sigma_g^2 + \sigma_r^2 + 2cov(g, r)$$

Because we established that the covariance between the genotypic value and the residual error is zero $cov(g, r) = 0$ then:

$$\sigma_y^2 = \sigma_g^2 + \sigma_r^2$$

Broad-sense heritability formula Substituting into the formula we get:

$$H^2_{plot} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_r^2}$$

Broad-sense heritability (on a plot basis) is the ratio of genetic variance to phenotypic variance and indicates the proportion of phenotypic variation attributable to genetic differences. Plot-based heritability is the heritability of a plot. This estimation can be useful when comparing efficiency of an experiment desing or to compare between experiments that have different number of replications. Breeders routinely include more than one plot per genotype, resulting in replicated plots for the same genotype. We can still think of heritability in this scenario of replicated genotypes. The next section deals with this notion, usually defined as the entry-mean heritability.

Entry-Mean Heritability

Entry-mean heritability H^2_{entry} refers to the heritability based on the average phenotype for each genotype when the genotypes are replicated in the experiment. For a completely randomized design (CRD) with balanced data, broad-sense heritability on an entry-mean basis is defined as

$$H^2_{entry} = \frac{\sigma^2_g}{\sigma^2_g + \bar{\sigma_r}^2}$$

Where:

- σ^2_g is the variance attributed to differences in genotypes in the experiment.
- $\bar{\sigma_r}^2$ is the variance of the **average** residual error (also known as prediction error variance). The term $\bar{\sigma_r}^2$ can be thought of as the variance in the different measurements (reps) for a given genotype in this CRD example.

What is the average residual error variance for the entry-mean heritability? The challenge in determining the appropriate heritability for a given experiment lies in accurately calculating the prediction error variance associated with the mean phenotype: $(\bar{\sigma_r}^2)$.The value of $\bar{\sigma_r}^2$ *depends on the experimental design.* In a completely randomized design, where genotypes are randomly assigned to plots within a field, it takes on a specific meaning. Let's take a look. The linear model used to represent a completely randomized design is:

$$y_{ij} = g_i + r_{ij}$$

Here, i represents the genotype identifier, and j represents the replicate identifier. Replicates are the multiple measurements on a given genotype. We have n total genotypes and m total replicates.

- y_{ij} denotes the response for each replication of every genotype.
- g_i stands for the genotype.
- r_{ij} represents the *residual error*, capturing differences between different replications of the same genotype. The j^{th} measurement for the i^{th} genotype, has an error associated with it. It follows that r_{ij} is a random variable normally distributed with mean zero and a variance parameter, expressed as $r_{ij} \sim N(0, \sigma^2_r)$.

Average residual error What is the average residual error for the i^{th} genotype? It is intuitive that for a given genotype, the average of the residual error is just the sum of all the errors divided by the number or replicates. For a CRD with m reps it looks like this:

$$\bar{r_i}. = \frac{r_{i1} + r_{i2} + \cdots + r_{im}}{m}$$

The expression above represents the average residual error among m replications for the i^{th} genotype.

Variance of the average residual error We can take the variance of the expression above:

$$\bar{\sigma_r}^2 = var(\bar{r_i}.)$$

We can substitute the average residual error in the variance operator :

$$\bar{\sigma_r}^2 = var\left(\frac{r_{i1} + r_{i2} + \cdots + r_{im}}{m}\right)$$

Variance scaling rule. This rule states that when a random variable is multiplied by a constant, the variance of that random variable gets multiplied by the square of the constant and comes out of the variance operator. Our random variables r_i are multiplied by one over the number of replicates $\frac{1}{m}$, which is a constant. Taking advantage of **variance scaling rule.** we have:

$$\bar{\sigma_r}^2 = \frac{1}{m^2}\left[var\left(r_{i1} + r_{i2} + \cdots + r_{im}\right)\right]$$

Additive law of variance This rule states that the variance of the sum of independent random variables is equal to the sum of their variances:

$$\bar{\sigma_r}^2 = \frac{1}{m^2}\left[var(r_{i1}) + var(r_{i2}) + \cdots + var(r_{im})\right]$$

Homogenous variance assumption We assume, for convenience that the variables r_{i1} to r_{im} are independent and have the same variance. Then our formula simplifies to:

$$\bar{\sigma_r}^2 = \frac{1}{m^2}\left[m\sigma_r^2\right]$$

$$= \frac{\sigma_r^2}{m}$$

Average residual error In a completely randomized design, the variance of the mean residual error ($\bar{\sigma_r}^2$) equals the variance component associated with plot residual error (σ_r^2) divided by the number of replicates (m). This relationship can be expressed as:

$$\bar{\sigma_r}^2 = \frac{\sigma_r^2}{m}$$

Entry-mean heritability formula We can subsitute this value in the formula for the entry-mean heritability and we get:

$$H^2_{entry} = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_r^2}{m}}$$

where, σ_g^2 is the variance attributed to differences in genotypes in the experiment, $\frac{\sigma_r^2}{m}$ is the residual error variance divided by m reps.

Average residual error for other experimental designs It's crucial to note that determining the correct heritability for an experiment lies on precisely calculating the variance of the average residual error associated with the mean phenotype, denoted as $\bar{\sigma_r}^2$. This term can vary in complexity and may involve numerous parameters, depending on the specifics of the experiment. The column Prediction Error Variance in the table below shows some of the forms that the variance of this average residual can take.

Model	Prediction Error	Prediction Error Variance
$y_{ij} = g_i + r_{ij}$	$r_i = \bar{r_i}.$	$\frac{\sigma_r^2}{m}$
$y_{ij} = g_i + b_j + r_{ij}$	$r_i = \bar{b}. + \bar{r_i}.$	$\frac{\sigma_r^2}{m}$
$y_{ijk} = g_i + E_j + gE_{ij} + r_{ijk}$	$r_i = \bar{E}. + g\bar{E}_i. + \bar{r_i}..$	$\frac{\sigma_{gE}^2}{e} + \frac{\sigma_r^2}{me}$

Model	Prediction Error	Prediction Error Variance
$y_{ijkl} =$ $g_i + L_j + Y_k + gL_{ij} +$ $gY_{ik} + gYL_{ijk} + r_{ijkl}$	$r_i = \bar{L}.+\bar{Y}.+g\bar{L}_i.+g\bar{Y}_i.+gY\bar{L}_i..+r_i\bar{...}$	$\frac{\sigma^2_{gL}}{l} + \frac{\sigma^2_{gY}}{y} + \frac{\sigma^2_{gLY}}{ly} + \frac{\sigma^2_r}{lym}$

In the table, m is the number of reps, e is the number of environments, l is the number of locations, y is the number of years. The models 1 to 4 are respectively: CRD, RCBD, CRD in a multi-environmental trial using the E model, and CRD multi-environemntal trial using the LY model. Remember that $e = ly$, in simple terms, environment is the location-year combination.

What can we learn fromt the table above? As the complexity of the experimental design model increases, so does the complexity of the prediction error variance. An essential insight from the analysis is that elements not indexed by the genotype (i) in the model do not contribute to the prediction error variance. However, interactions involving genotype with factors like environment, location, or year do contribute to this variance. As shown above and because of how the math works out, the trend is that genotype by environment variance is divided by the number of environments, the genotype by location variance is divided by the number of locations, the genotype by location by year variance is divided by the product of the number locations and years and so forth.

Conclusion

It is critical to understand the type of experiment you have to estimate accurate heritabilities. Please note that this formulas work in the context of balanced designs which in reality tends to be the exception and not the rule. Breeders routinely use them in unbalanced designs at the cost of decreased precision, which in cases may be negligible or pronounced. Plant breeding selection experiments tend to be unbalanced by nature. We will see in the next handout, methods to estimate heritablity which are precise when used in unbalanced designs.

References

Lecture Notes 812 Selection Theory by Prof. Jeffrey Endelman and Prof. Natalia de Leon.

Variance of a Random variable