Heritability

Nuno Carvalho

April 16, 2025

1 SNP Heritability

These derivations are based on the Methods of [Yang et al., 2010].

1.1 Phenotype model

We can define a quantitative phenotype y as:

$$\mathbf{y} = \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{y}$: phenotypes. $N \times 1$ vector. Centered so that $\mathrm{E}[\mathbf{y}] = 0$.
  - N: number of samples.
- $\mathbf{X}_c$: normalized genotypes for causal variants. $N \times M_c$ matrix.
  - M_c: number of causal variants.
  - Normalized according to $\mathbf{X}_{c,i} = \frac{\mathbf{X}'_{c,i} - 2f_i}{\sqrt{2f_i(1-f_i)}}$.
    - $\mathbf{X}'_c$: allele dosages, taking on values of 0, 1, 2.
    - f_i: true population allele frequency for variant i.
    - Such that for each row (variant), $\mathrm{E}[\mathbf{X}_{c,i}] = 0$ and $\mathrm{Var}[\mathbf{X}_{c,i}] = 1$.
- $\boldsymbol{\beta}$: per-normalized-genotype causal effects. $M_c \times 1$ vector.
  - Assume infinitesimal model.
  - Drawn from $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2_\beta)$.
    - $\mathbf{I}$: $M_c \times M_c$ identity matrix.
    - σ^2_β: variance of causal effects.
- $\boldsymbol{\epsilon}$: residual effects (i.e. error or noise term). $N \times 1$ vector.
  - Drawn from $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}\sigma^2_\epsilon)$.
    - $\mathbf{I}$: $N \times N$ identity matrix.
    - σ^2_ϵ: residual variance.

We assume $\mathbf{X}_c$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ are all independent from each other. We can define the genetic effects as a single term, $\mathbf{g} = \mathbf{X}_c\boldsymbol{\beta}$, meaning that:

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon} \quad \text{where} \quad \mathbf{g} \sim \mathcal{N}(0, I\sigma^2_g) \quad \text{where} \quad \sigma^2_g = M_c\sigma^2_\beta$$

We interpret σ^2_g as variance of total additive genetic effects on the phenotype.

1.2 Variance of the phenotype

By making use of the independence between terms, we can define the variance-covariance matrix of $\mathbf{y}$ as:

$$\begin{aligned}
\text{Var}[\mathbf{y}] &= \text{Var}[\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\
&= \text{Var}[\mathbf{X}_c]\text{Var}[\boldsymbol{\beta}] + \text{Var}[\boldsymbol{\epsilon}] \\
&= (\mathbf{X}_c\mathbf{X}_c^\intercal)\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2 \\
&= (\mathbf{X}_c\mathbf{X}_c^\intercal)\frac{\sigma_g^2}{M_c} + \mathbf{I}\sigma_\epsilon^2 \\
&= \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2
\end{aligned}$$

Where we define $\mathbf{G} = \frac{\mathbf{X}_c\mathbf{X}_c^\intercal}{M_c}$ as the $N \times N$ genetic relationship matrix (GRM) between individuals. The G_{ii} element is the variance of individual i's normalized genotype vector, while the G_{ij} element is the covariance of individuals i and j's normalized genotype vectors.

Narrow-sense heritability is defined as the proportion of phenotypic variance, σ_P^2, explained by additive genetic effects:

$$h^2 = \frac{\sigma_g^2}{\sigma_P^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$$

1.3 Estimating the GRM

In practice, we likely do not know the exact set of causal variants and instead must estimate the GRM using a set of genotyped SNPs:

$$\mathbf{A} = \frac{\mathbf{X}\mathbf{X}^\intercal}{M}$$

Where $\mathbf{A}$ is the estimated GRM, X is the normalized genotype matrix of our genotyped SNPs, and M is the number of genotyped SNPs. Note that because we are also working with a sample, X is normalized using sample allele frequencies, $\mathbf{p}$:

$$\mathbf{X}_i = \frac{\mathbf{X}_i' - 2p_i}{\sqrt{2p_i(1-p_i)}}$$

However, this equation for A ignores the sampling error associated with each SNP. Let's consider the covariance computation between two individuals for SNP i, which is then

summed across M SNPs to get the value for A_{jk}. When $j \neq k$:

$$\begin{aligned} A_{ijk} &= x_{ij}x_{ik} \\ &= \frac{x'_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}} \frac{x'_{ik} - 2p_i}{\sqrt{2p_i(1-p_i)}} \\ &= \frac{(x'_{ij} - 2p_i)(x'_{ik} - 2p_i)}{2p_i(1-p_i)} \end{aligned}$$

Because x'_{ij} and x'_{ik} are independent from each other and p_i is a constant:

$$\begin{aligned} \mathrm{Var}[A_{ijk}] &= \mathrm{Var}[\frac{(x'_{ij} - 2p_i)(x'_{ik} - 2p_i)}{2p_i(1-p_i)}] \\ &= \frac{\mathrm{Var}[(x'_{ij} - 2p_i)]\mathrm{Var}[(x'_{ik} - 2p_i)]}{(2p_i(1-p_i))^2} \\ &= \frac{\mathrm{Var}[(x'_{ij})]\mathrm{Var}[(x'_{ik})]}{(2p_i(1-p_i))^2} \\ &= \frac{(2p_i(2-p_i))(2p_i(2-p_i))}{(2p_i(1-p_i))^2} \\ &= 1 \end{aligned}$$

So, the variance in A_{jk} is independent of allele frequency. But this is not the case when $j = k$:

$$\begin{aligned} A_{ijj} &= x_{ij}^2 \\ &= (\frac{x'_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}})^2 \\ &= \frac{(x'_{ij} - 2p_i)^2}{2p_i(1-p_i)} \end{aligned}$$

For simplicity, let's denote $Z = 2p_i(1-p_i)$ and make use of $\mathrm{Var}[Y] = \mathrm{E}[Y^2] - \mathrm{E}[Y]^2$:

$$\begin{aligned} \mathrm{Var}[A_{ijj}] &= \mathrm{Var}[\frac{(x'_{ij} - 2p_i)^2}{Z}] \\ &= \frac{\mathrm{Var}[(x'_{ij} - 2p_i)^2]}{(Z)^2} \\ &= \frac{\mathrm{E}[((x'_{ij} - 2p_i)^2)^2] - \mathrm{E}[(x'_{ij} - 2p_i)^2)]^2}{(Z)^2} \\ &= \frac{(Z) - (Z)^2}{(Z)^2} \\ &= \frac{(Z)(1-Z)}{(Z)^2} \\ &= \frac{1-Z}{Z} \\ &= \frac{1 - 2p_i(1-p_i)}{2p_i(1-p_i)} \end{aligned}$$

The full derivation for why $\mathrm{E}[((x'_{ij} - 2p_i)^2)^2] = \mathrm{E}[(x'_{ij} - 2p_i)^2] = 2p_i(1-p_i)$ is very lengthy algebraically, but can be shortcutted by using the formula for the higher moments of a binomially distributed variable, where $n = 2$ and $p = p_i$. Importantly, the variance of A_{jj} therefore depends on the allele frequencies of the SNPs, even after normalization.