# Heritability

Narrow-sense heritability ($h^2$) is a population parameter that describes the proportion of phenotypic variation that attributable to *additive* genetic variation [7]. Initially, heritability was estimated by means of pedigree analysis. Later, methods were developed to estimate heritability within genome-wide association studies (GWAS). Each approach has advantages and limitations. Pedigree-based heritability $h^2_{\text{ped}}$ is susceptible to inflation due to shared environmental and non-additive genetic effects. GWAS or SNP-based heritability $h^2_{\text{SNP}}$ is less susceptible to inflation because the variation explained is estimated using conventionally-unrelated individuals who are unlikely to consistently share a common environment or non-additive genetic effects. However, the value of $h^2_{\text{SNP}}$ depends on which SNPs are included in the analysis. Because a GWAS typically will not contain all causal variants for a trait, $h^2_{\text{SNP}} \leq h^2$ in general.

# Quantitative Trait SNP-Heritability

SNP-based heritability is estimated via a linear mixed effects model of the form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{2.1}$$

- $\boldsymbol{y}$ is an $n \times 1$ phenotype vector.
- $\boldsymbol{X}$ is an $n \times p$ covariate matrix, not including genotype, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects.
- $\boldsymbol{W}$ is an $n \times m$ standardized genotype matrix, and $\boldsymbol{u} \sim N(0, \sigma^2_u \boldsymbol{I})$ is a vector of random genetic effects. $W_{ij} = (G_{ij} - p_j)/\sqrt{2p_j(1-p_j)}$ where $G_{ij} \in \{0, 1, 2\}$ is the number of minor alleles for subject $i$ at SNP $j$, and $p_j$ is a minor allele frequency of SNP $j$. Because $\boldsymbol{u}$ is constrained to follow a normal distribution with a single free parameter ($\sigma^2_u$), (2.1) is estimable even when $m \gg n$.
- $\boldsymbol{\epsilon} \sim N(0, \sigma^2_\epsilon)$ is an $n \times 1$ residual vector.

Model (2.1) can be recast as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{g} \sim N(0, \sigma^2_g \boldsymbol{A})$ where $\boldsymbol{A} = m^{-1}\boldsymbol{W}\boldsymbol{W}'$ and $\sigma^2_g = m\sigma^2_u$. $\boldsymbol{A}$ is the empirical genetic relatedness matrix (GRM), $\sigma^2_g$ is the total genetic variance, and $\sigma^2_u$ is the per-SNP genetic variance. The total phenotypic variance is:

$$\mathbb{V}(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2_g \boldsymbol{A} + \sigma^2_\epsilon \boldsymbol{I}$$

The **SNP-heritability** is:

$$h^2_{\text{SNP}} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2} = \frac{m\sigma_u^2}{m\sigma_u^2 + \sigma_\epsilon^2}.$$

## 2.1 REML Estimation

The variance components $\sigma_g^2$ and $\sigma_\epsilon^2$ from (2.2) may be estimated via restricted maximum likelihood. In the single component case, the restricted log likelihood is:

$$\ell(\sigma_g^2, \sigma_\epsilon^2) = -\frac{1}{2}\big\{\ln\det(\boldsymbol{V}) + \ln\det(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}) + \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}\big\},$$

where $\boldsymbol{V} = \sigma_g^2\boldsymbol{A} + \sigma_\epsilon^2\boldsymbol{I}$ and:

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}.$$

The single component model in (2.2), introduced by Yang *et al* [2], is sensitive how the minor allele frequency (MAF) and linkage disequilibrium (LD) of causal variants compare with those of tagging variants. Extending the model to utilize multiple genetic components makes the estimated heritability robust to these properties [5]. The multi-component model takes the form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{k=1}^{K}\boldsymbol{g}_k + \boldsymbol{\epsilon},$$

where $\boldsymbol{g}_k \sim N(0, \sigma_{gk}^2\boldsymbol{A}_k)$ and $\boldsymbol{A}_k$ is the GRM calculated using only those variants in the $k$th stratum, where strata are defined based on MAF and LD. The SNP-heritability under a multi-component model is:

$$h^2_{\text{SNP}} = \frac{\sum_{k=1}^{K}\sigma_k^2}{\sum_{k=1}^{K}\sigma_k^2 + \sigma_\epsilon^2}.$$

## 2.2 Haseman-Elston Regression

Consider the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\zeta}, \qquad \boldsymbol{\zeta} = \boldsymbol{g} + \boldsymbol{\epsilon},$$

where $\boldsymbol{g} \sim N(0, \sigma_g^2\boldsymbol{A})$ and $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2\boldsymbol{I})$. Observe that $\mathbb{E}(\boldsymbol{\zeta}) = 0$ and that:

$$\mathbb{E}(\zeta_i\zeta_j) = \sigma_g^2A_{ij} + \sigma_\epsilon^2\delta_{ij} = (A_{ij}, \delta_{ij})'(\sigma_g^2, \sigma_\epsilon^2).$$

To estimate the variance components $\boldsymbol{\theta} = (\sigma_g^2, \sigma_\epsilon^2)'$ construct an outcome $\boldsymbol{z} = (\zeta_i\zeta_j)$ for $i \le j$ and a corresponding design matrix $\boldsymbol{B} = (A_{ij}, \delta_{ij})$. Although $\zeta_i$ is unobserved, it can be replaced by the consistent estimator $\hat{\zeta}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares (OLS) estimator from regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ [6]. Regression of $\hat{z} = (\hat{\zeta}_i\hat{\zeta}_j)$ on $\boldsymbol{B}$ provides a method-of-moments estimator for $\boldsymbol{\theta}$.

### 2.3 Inference

Given an estimate of $\boldsymbol{\theta} = (\sigma_g^2, \sigma_\epsilon^2)'$, obtained (e.g.) by REML or Haseman-Elston regression, and an estimate $\boldsymbol{C}_{\theta\theta'}$ of its covariance, the SNP-heritability and its standard error may be calculated as follows. Let $g(u, v) = u/(u + v)$, such that $h^2_{\text{SNP}} = g(\boldsymbol{\theta})$. The gradient of $g$ is:

$$\nabla g = \begin{pmatrix} v/(u+v)^2 \\ -u/(u+v)^2 \end{pmatrix}$$

By the $\Delta$-method:

$$\hat{h}^2_{\text{SNP}} \dot{\sim} N\big\{h^2_{\text{SNP}}, \nabla g(\boldsymbol{\theta})'\boldsymbol{C}_{\theta\theta'}\nabla g(\boldsymbol{\theta})\big\}.$$

### 2.4 LD Score Regression

LD Score Regression [4] allows for estimation of heritability using summary statistics. The fundamental equations of LDSC is:

$$\mathbb{E}(\chi_j^2 | \ell_j) = (na + 1) + \frac{n h^2_{\text{SNP}}}{m}\ell_j,$$

- $n$ is the sample size and $m$ the number of SNPs.
- The **LD score** $\ell_j = \sum_k r_{jk}^2$ is the sum of the squared correlation of the $j$th SNP with all other SNPs.
- $a$ is a measure of inflation attributed to confounding biases.

LDSC tends to underestimate heritability, particularly when the causal variants are rare, but is a robust method of estimating a lower bound [8].

# Binary Trait SNP-Heritability

### 3.1 Liability Threshold Model

Under the liability threshold model, an individual's latent disease liability is:

$$L_i = G_i + \epsilon_i,$$

where $G_i \sim N(0, h_L^2)$, $\epsilon_i \sim N(0, 1 - h_L^2)$, and $h_L^2$ is the liability-scale heritability. An individual is a case ($Y_i = 1$) if their liability exceeds a threshold $\tau$ determined by the prevalence $K$ of the disease in the population:

$$K = \mathbb{P}(Y = 1) = \mathbb{P}(L > \tau) = 1 - \Phi(\tau). \tag{3.1}$$

Because the liability $L_i$ is not directly observed, the observed-scale heritability $h^2_{\text{obs}}$ is first estimated, based on $Y_i \in \{0, 1\}$, then transformed to liability-scale heritability using the prevalence $K$ and ascertainment proportion $P$ (i.e. the proportion of subjects who are cases due to the sampling scheme). Lee *et al* [1] developed a formula for converting observed-scale heritability to liability-scale heritability:

$$h^2_L = h^2_{\text{obs}} \frac{K(1-K)}{\phi^2(\tau)} \cdot \frac{K(1-K)}{P(1-P)},$$

where $\tau$ is the liability threshold (3.1) and $\phi$ is the standard normal PDF. Golan *et al* [3] report that estimating the observed-scale heritability for a binary trait by means of REML (2.1) introduces substantial bias. They propose estimating $h^2_{\text{obs}}$ by means of phenotype-correlation genotype-correlation (PCGC), which extends Haseman-Elston regression to the binary setting and enables covariate correction. PCGC was further developed by Weissbrod *et al* [9].

# References

[1] SH Lee et al. "Estimating Missing Heritability for Disease from Genome-wide Association Studies". In: *Am J Hum Genet* 88.3 (2011), pp. 294–305. DOI: 10.1016/j.ajhg.2011.02.002.

[2] Yang, J and Lee, SH and Goddard, ME and Visscher, PM. "GCTA: A Tool for Genome-wide Complex Trait Analysis". In: *Am J Hum Genet* 88.1 (2011), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011.

[3] D Golan, ES Lander, and S Rosset. "Measuring missing heritability: inferring the contribution of common variants". In: *Proc Natl Acad Sci U S A* 111.49 (2014), E5272–81. DOI: 10.1073/pnas.1419064111.

[4] BK Bulik-Sullivan, PH Loh, HK Finucane, et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nat Genet* 47.3 (2015), pp. 291–295. DOI: 10.1038/ng.3211.

[5] J Yang, A Bakshi, Z Zhu, et al. "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index". In: *Nat Genet* 47.10 (2015), pp. 1114–1120. DOI: 10.1038/ng.3390.

[6] T Sofer. "Confidence intervals for heritability via Haseman-Elston regression". In: *Stat Appl Genet Mol Biol* 16.4 (2017), pp. 259–273. DOI: 10.1515/sagmb-2016-0076.

[7] Yang, J and Zeng, J and Goddard, ME and Wray, NR and Visscher, PM. "Concepts, estimation and interpretation of SNP-based heritability". In: *Nat Genet* 49.9 (2017), pp. 1304–1310. DOI: 10.1038/ng.3941.

[8] LM Evans, R Tahmasbi, SI Vrieze, et al. "Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits". In: *Nat Genet* 50.5 (2018), pp. 737–745. DOI: 10.1038/s41588-018-0108-x.

[9] O Weissbrod, J Flint, and S Rosset. "Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics". In: *Am J Hum Genet* 103.1 (2018), pp. 89–99. DOI: 10.1016/j.ajhg.2018.06.002.