The American Journal of Human Genetics, Volume *109*

Supplemental information

Liability-scale heritability estimation for biobank studies of low-prevalence disease

Sven E. Ojavee, Zoltan Kutalik, and Matthew R. Robinson

Supplementary Information

Derivation of the error variance term

Suppose that the genetic value (of an individual) g and error term e are from normal distributions $g \sim N(0, h_l^2)$ and $e \sim N(0, 1-h_l^2)$, where h_l^2 is the underlying liability scale heritability. Then the underlying liability $l = g + e$ and the binary trait y with a prevalence of K is defined as

$$y = \begin{cases} 0, & \text{if } l < \Phi^{-1}(1-K) \\ 1, & \text{if } l \geq \Phi^{-1}(1-K) \end{cases} \tag{1}$$

From this we will derive the error variance term $E(Var(y|c+zg))$ where c is some constant and z is the standard Gaussian density evaluated at $\Phi^{-1}(1-K)$ as shown in [?]. As $c+zg$ is a linear combination of g we can equivalently find $E(Var(y|g))$. First, we note the conditional distribution of y given g

y	0	1
$P(y\|g)$	$P(\frac{e}{\sqrt{1-h_l^2}} < \frac{\Phi^{-1}(1-K)-g}{\sqrt{1-h_l^2}}) = \Phi(\frac{\Phi^{-1}(1-K)-g}{\sqrt{1-h_l^2}})$	$P(\frac{e}{\sqrt{1-h_l^2}} \geq \frac{\Phi^{-1}(1-K)-g}{\sqrt{1-h_l^2}}) = \Phi(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}})$

(2)

As y can be equal to only 0 or 1, we can write

$$Var(y|g) = E(y^2|g) - E(y|g)^2 = E(y|g) - E(y|g)^2 = P(y=1|g) - P(y=1|g)^2 = \Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big) - \Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)^2. \tag{3}$$

To find $E(Var(y|g))$ we need to find $E(\Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big))$ and $E(\Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)^2)$. For this, we use auxiliary standardised Gaussian random variables X, X_1 and X_2 that are independent of g and X_1 is independent of X_2. From this it follows that $Var(X\sqrt{1-h_l^2} - g) = 1$ and using the law of total probability we get

$$E(\Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)) = P\Big(X \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big) = P(X\sqrt{1-h_l^2} - g \leq -\Phi^{-1}(1-K)) = \Phi(-\Phi^{-1}(1-K)) = 1 - \Phi(\Phi^{-1}(1-K)) = K. \tag{4}$$

Secondly, we see that we can analogously use X_1 and X_2 to find the second moment of $\Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)$. For this we need to find the following correlation

$$cor(X_1\sqrt{1-h_l^2} - g, X_2\sqrt{1-h_l^2} - g) = E((X_1\sqrt{1-h_l^2} - g)(X_2\sqrt{1-h_l^2} - g)) = E(g^2) = h_l^2. \tag{5}$$

Now we express the expectation using a cumulative distribution function of a bivariate Gaussian distribution of two random variables that have a correlation of h_l^2

$$E(\Phi\Big(\frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)^2) = E(P\Big(X_1 \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)P\Big(X_2 \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)) =$$
$$E(P\Big(X_1 \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}, X_2 \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)) = P\Big(X_1 \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}, X_2 \leq \frac{g-\Phi^{-1}(1-K)}{\sqrt{1-h_l^2}}\Big)$$
$$P(X_1\sqrt{1-h_l^2} - g \leq -\Phi^{-1}(1-K), X_2\sqrt{1-h_l^2} - g \leq -\Phi^{-1}(1-K)) =$$
$$\tilde{\Phi}(-\Phi^{-1}(1-K), -\Phi^{-1}(1-K), h_l^2) = \tilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h_l^2), \quad (6)$$

where $\tilde{\Phi}(x_1, x_2, \rho)$ is the cumulative distribution function of a standardised bivariate Gaussian distribution with a correlation of ρ. The first equation follows from the definition of cumulative distribution function, second from the independence of X_1 and X_2, third from the law of total probability. Thus, by combining the two last results, we get the final expression for the error variance

$$E(Var(y|c+zg)) = K - \tilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h_l^2). \quad (7)$$