Correlation between Z-scores in a GWAS meta-analysis

Anna Hutchinson

January 2021

1 The problem

Assume that two separate genome-wide association studies (GWAS) have been performed on two SNPs: SNP A and SNP B.

In study 1, estimated effect sizes, β_{A1} and β_{B1} and their corresponding standard errors, se_{A1} and se_{B1} are generated for SNPs A and B respectively. Similarly, in study 2 estimated effect sizes, β_{A2} and β_{B2} and their corresponding standard errors, se_{A2} and se_{B2} are generated.

An overall measure of strength against the null for each SNP is then calculated using fixed effect meta analysis, generating Z_A and Z_B.

Question: What is the expected correlation between the meta analysis Z_A and Z_B?

2 Details

2.1 Fixed effect meta-analysis

The fixed effect (FE) method assumes that the magnitude of the true effect is common or fixed in every study in the meta-analysis [2]. The inverse-variance-weighted effect-size method [1] is a popular choice for this. In this method, each estimated β_i is given by

$$\beta_i = \mu + \epsilon_i \tag{1}$$

where μ is the true common effect and ϵ_i is the within-study error.

In our example, this would be

$$\begin{pmatrix}\beta_{A1}\\ \beta_{A2}\end{pmatrix} = \mu_A + \begin{pmatrix}\epsilon_1\\ \epsilon_2\end{pmatrix} \tag{2}$$

and

$$\begin{pmatrix}\beta_{B1}\\ \beta_{B2}\end{pmatrix} = \mu_B + \begin{pmatrix}\epsilon_1\\ \epsilon_2\end{pmatrix} \tag{3}$$

where μ_A and μ_B are the true common effects of SNPs A and B and ϵ_1 and ϵ_2 are the within-study errors for study 1 and 2.

If sample sizes of the studies are sufficiently large, β_i is normally distributed with variance $(se_i)^2$. In inverse-variance weighting, we weight by the inverse of the variance: $W_i = (se_i^2)^{-1}$. The inverse-variance-weighted effect-size estimator is then the sum of β_i with weights W_i so that

$$\beta_{FE} = \frac{\sum W_i \beta_i}{\sum W_i}. \tag{4}$$

The variance of β_{FE} is $V_{FE} = \frac{1}{\sum W_i}$ so that the standard error is $SE(\beta_{FE}) = (\sum W_i)^{-1/2}$.

The summary score is then given by

$$Z_{FE} = \frac{\beta_{FE}}{SE(\beta_{FE})} = \frac{\sum W_i \beta_i}{\sqrt{\sum W_i}} \tag{5}$$

which follows $N(0,1)$ under the null hypothesis of no association, so that the p-value can be calculated as $p_{FE} = 2\Phi(-|Z_{FE}|)$.

In our example the meta-analysed Z scores for SNPs A and B are:

$$Z_A = \frac{W_{A1}\beta_{A1} + W_{A2}\beta_{A2}}{\sqrt{W_{A1} + W_{A2}}} \tag{6}$$

and

$$Z_B = \frac{W_{B1}\beta_{B1} + W_{B2}\beta_{B2}}{\sqrt{W_{B1} + W_{B2}}}. \tag{7}$$

Note that we can also write this as

$$Z_A = \frac{\frac{\beta_{A1}}{(se_{A1})^2} + \frac{\beta_{A2}}{(se_{A2})^2}}{\sqrt{\frac{1}{(se_{A1})^2} + \frac{1}{(se_{A2})^2}}}. \tag{8}$$

3 Solution

The correlation of the z-scores is the covariance of the z-scores since they are $N(0,1)$. Thus,

$$\begin{aligned} cor\left(\frac{\beta_{A1}}{se_{A1}}, \frac{\beta_{B1}}{se_{B1}}\right) &= cov\left(\frac{\beta_{A1}}{se_{A1}}, \frac{\beta_{B1}}{se_{B1}}\right) = E\left(\frac{\beta_{A1}}{se_{A1}} \times \frac{\beta_{B1}}{se_{B1}}\right) = r_1 \\ cor\left(\frac{\beta_{A2}}{se_{A2}}, \frac{\beta_{B2}}{se_{B2}}\right) &= cov\left(\frac{\beta_{A2}}{se_{A2}}, \frac{\beta_{B2}}{se_{B2}}\right) = E\left(\frac{\beta_{A2}}{se_{A2}} \times \frac{\beta_{B2}}{se_{B2}}\right) = r_2 \end{aligned} \tag{9}$$

And

$$cor(Z_A, Z_B) = cov(Z_A, Z_B) = E(Z_A \times Z_B). \tag{10}$$

So,

$$E(Z_A \times Z_B) = E\left(\frac{\frac{\beta_{A1}}{(se_{A1})^2} + \frac{\beta_{A2}}{(se_{A2})^2}}{\sqrt{\frac{1}{(se_{A1})^2} + \frac{1}{(se_{A2})^2}}} \times \frac{\frac{\beta_{B1}}{(se_{B1})^2} + \frac{\beta_{B2}}{(se_{B2})^2}}{\sqrt{\frac{1}{(se_{B1})^2} + \frac{1}{(se_{B2})^2}}} \right). \tag{11}$$

I focus on expanding the quantity inside the expectation, rewriting $(se_X)^2$ as V_X.

The numerator in the expectation of Equation (11) is:

$$\begin{aligned} &= \left(\frac{\beta_{A1}}{V_{A1}} + \frac{\beta_{A2}}{V_{A2}}\right)\left(\frac{\beta_{B1}}{V_{B1}} + \frac{\beta_{B2}}{V_{B2}}\right) \\ &= \left(\frac{\beta_{A1}V_{A2} + \beta_{A2}V_{A1}}{V_{A1}V_{A2}}\right)\left(\frac{\beta_{B1}V_{B2} + \beta_{B2}V_{B1}}{V_{B1}V_{B2}}\right) \\ &= \frac{\beta_{A1}V_{A2}\beta_{B1}V_{B2} + \beta_{A2}V_{A1}\beta_{B1}V_{B2} + \beta_{A1}V_{A2}\beta_{B2}V_{B1} + \beta_{A2}V_{A1}\beta_{B2}V_{B1}}{V_{A1}V_{A2}V_{B1}V_{B2}} \end{aligned} \tag{12}$$

The denominator is:

$$\begin{aligned} &= \left(\sqrt{\frac{1}{V_{A1}} + \frac{1}{V_{A2}}}\right)\left(\sqrt{\frac{1}{V_{B1}} + \frac{1}{V_{B2}}}\right) \\ &= \sqrt{\left(\frac{1}{V_{A1}} + \frac{1}{V_{A2}}\right)\left(\frac{1}{V_{B1}} + \frac{1}{V_{B2}}\right)} \\ &= \sqrt{\left(\frac{V_{A2} + V_{A1}}{V_{A1}V_{A2}}\right)\left(\frac{V_{B2} + V_{B1}}{V_{B1}V_{B2}}\right)} \\ &= \sqrt{\frac{V_{A2}V_{B2} + V_{A1}V_{B2} + V_{A2}V_{B1} + V_{A1}V_{B1}}{V_{A1}V_{A2}V_{B1}V_{B2}}} \end{aligned} \tag{13}$$

Putting this together (still working inside the expectation for now),

$$
\begin{aligned}
Z_A \times Z_B = & \frac{\beta_{A1}V_{A2}\beta_{B1}V_{B2} + \beta_{A2}V_{A1}\beta_{B1}V_{B2} + \beta_{A1}V_{A2}\beta_{B2}V_{B1} + \beta_{A2}V_{A1}\beta_{B2}V_{B1}}{V_{A1}V_{A2}V_{B1}V_{B2}} \\
& \times \sqrt{\frac{V_{A1}V_{A2}V_{B1}V_{B2}}{V_{A2}V_{B2} + V_{A1}V_{B2} + V_{A2}V_{B1} + V_{A1}V_{B1}}} \\
= & (\frac{\beta_{A1}\beta_{B1}}{se_{A1}se_{B1}} \times se_{A2}se_{B2} + \frac{\beta_{A2}\beta_{B1}}{se_{A2}se_{B1}} \times se_{A1}se_{B2} \\
& + \frac{\beta_{A1}\beta_{B2}}{se_{A1}se_{B2}} \times se_{A2}se_{B1} + \frac{\beta_{A2}\beta_{B2}}{se_{A2}se_{B2}} \times se_{A1}se_{B1}) \\
& \times \frac{1}{se_{A2}se_{B2} + se_{A1}se_{B2} + se_{A2}se_{B1} + se_{A1}se_{B1}}
\end{aligned} \tag{14}
$$

When taking expectations in order to derive our final quantity of interest (see Equation 11), I think that we are able to drop the $\frac{\beta_{A2}\beta_{B1}}{se_{A2}se_{B1}}$ and $\frac{\beta_{A1}\beta_{B2}}{se_{A1}se_{B2}}$ terms since these are $cor(Z_{A2}, Z_{B1})$ and $cor(Z_{A1}, Z_{B2})$ respectively (i.e. correlations between different SNP Z-scores in different studies)... perhaps not but for now I'm going to drop them and see what happens (can always just use their actual values which we know from the seperate GWASs?).

Also, recall from equation (9) that $r_1 = E(\frac{\beta_{A1}}{se_{A1}} \times \frac{\beta_{B1}}{se_{B1}}) = cor(\frac{\beta_{A1}}{se_{A1}}, \frac{\beta_{B1}}{se_{B1}})$ and that $r_2 = E(\frac{\beta_{A2}}{se_{A2}} \times \frac{\beta_{B2}}{se_{B2}}) = cor(\frac{\beta_{A2}}{se_{A2}}, \frac{\beta_{B2}}{se_{B2}})$.

We can therefore write (assuming lots of things are independent so that $E(XY) = E(X)E(Y)$):

$$
\begin{aligned}
E(Z_A \times Z_B) = & \left(E\left(\frac{\beta_{A1}\beta_{B1}}{se_{A1}se_{B1}}\right) \times E(se_{A2}se_{B2}) + E\left(\frac{\beta_{A2}\beta_{B2}}{se_{A2}se_{B2}}\right) \times E(se_{A1}se_{B1})\right) \\
& \times E\left(\frac{1}{se_{A2}se_{B2} + se_{A1}se_{B2} + se_{A2}se_{B1} + se_{A1}se_{B1}}\right) \\
= & (r_1 \times E(se_{A2}se_{B2}) + r_2 \times E(se_{A1}se_{B1})) \\
& \times E\left(\frac{1}{se_{A2}se_{B2} + se_{A1}se_{B2} + se_{A2}se_{B1} + se_{A1}se_{B1}}\right)
\end{aligned} \tag{15}
$$

I think that we're able to remove the expectations from the se only terms, since these are constants. These values can be calculated using study specific quantities. I.e.

$$
se_i = \sqrt{\frac{1}{2 \times N \times MAF_i \times (1 - MAF_i) \times s \times (1 - s)}} \tag{16}
$$

where N and s are specific quantities for each study.

References

[1] Paul I.W. de Bakker, Manuel A.R. Ferreira, Xiaoming Jia, Benjamin M. Neale, Soumya Raychaudhuri, and Benjamin F. Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, 17(R2):R122–R128, October 2008.

[2] C H Lee, E Eskin, and B Han. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics*, 33(14):i379–i388, July 2017.