FastMap: Fast Association Mapping in Inbred Mouse Populations Supplementary Materials

Daniel M Gatti[1]*, Andrey A Shabalin [2]*, Tieu-Chong Lam[1], Fred A. Wright[3], Ivan Rusyn[1]†, and Andrew B Nobel[2,3]†

October 1, 2008

[1]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, US
[2]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, US
[3]Department of Biostatistics, University of North Carolina at Chapel Hill, US

Equivalence of different test statistics

In this section we establish the equivalence of several test statistics (r^2, t^2, F, and LR) commonly used to test the associations between a binary vector, resulting from a homozygous SNP, and a real-valued vector, resulting from measurements of gene expression. In particular, we note that

$$F = t^2 = \frac{r^2}{1-r^2}(n-2) \quad \text{and} \quad LR = -\log(1-r^2)$$

Notation

Let $S = (0, 0, ..., 0, 1, 1, ..., 1)$ be a binary vector of length n, with n_0 zeros and n_1 ones, and let $G = (g_1, g_2, ..., g_n)$ be an n-vector with real-valued components. All four statistics we will consider are scale invariant, so without

*equally contributing coauthors
†to whom correspondence should be addressed

loss of generality we can assume that G is standardized:

$$\sum_{i=1}^{n} g_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} g_i^2 = 1.$$

Denote by $\bar{g}_0$ and $\bar{g}_1$ the average of G over those samples where S equal to 0 and 1, respectively:

$$\bar{g}_0 = \sum_{i=1}^{n_0} g_i/n_0 \quad \text{and} \quad \bar{g}_1 = \sum_{i=n_0+1}^{n} g_i/n_1.$$

It follows from $\sum_{i=1}^{n} g_i = 0$ that $n_0\bar{g}_0 = -n_1\bar{g}_1$, and in particular, $n_0^2\bar{g}_0^2 = n_1^2\bar{g}_1^2$.

Correlation

The correlation of G and S is defined as

$$r = \frac{\mathrm{cov}(G,S)}{\sqrt{\mathrm{var}(S)\mathrm{var}(G)}}.$$

The numerator is equal to

$$\mathrm{cov}(G,S) = \frac{1}{n-1}\sum_{i=1}^{n} s_i g_i = \frac{1}{n-1}\sum_{s_i=1} g_i = \bar{g}_1 n_1/(n-1).$$

The first term in the denominator simplifies to

$$\mathrm{var}(S) = \frac{1}{n-1}\left[\sum s_i^2 - \frac{1}{n}(\sum s_i)^2\right] = \frac{1}{n-1}\left[n_1 - \frac{n_1^2}{n}\right] = \frac{n_1(n-n_1)}{(n-1)n} = \frac{n_0 n_1}{(n-1)n},$$

and the second is equal to

$$\mathrm{var}(G) = \frac{1}{n-1}\left[\sum g_i^2 - \frac{1}{n}(\sum g_i)^2\right] = \frac{1}{n-1}.$$

Combining the previous three expressions, we find that

$$r = \frac{\bar{g}_1 n_1}{\sqrt{n_0 n_1/n}} = \bar{g}_1\sqrt{\frac{n_1}{n_0}n}.$$

T-statistic

The t-statistic for difference between $\bar{g}_0$ and $\bar{g}_1$ is defined as

$$t = \frac{\bar{g}_1 - \bar{g}_0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}},$$

where the estimate of the noise variance is

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n_0}(g_i - \bar{g}_0)^2 + \sum_{i=n_0+1}^{n}(g_i - \bar{g}_1)^2\right)$$

Simplifying $\hat{\sigma}^2$ yields

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n} g_i^2 - n_0\bar{g}_0^2 - n_1\bar{g}_1^2\right)$$

It then follows from the equations $\sum_{i=1}^{n} g_i^2 = 1$ and $n_0^2\bar{g}_0^2 = n_1^2\bar{g}_1^2$ that

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(1 - \frac{n_1(n_0+n_1)}{n_0}\bar{g}_1^2\right) = \frac{1}{n-2}\left(1 - \bar{g}_1^2\frac{n_1}{n_0}n\right) = \frac{1-r^2}{n-2}$$

Note that

$$\sqrt{\frac{1}{n_1} + \frac{1}{n_0}} = \sqrt{\frac{n_0+n_1}{n_0 n_1}} = \sqrt{\frac{n}{n_0 n_1}}$$

and that the numerator of t can be rewritten as

$$\bar{g}_1 - \bar{g}_0 = (1 + \frac{n_1}{n_0})\bar{g}_1 = \frac{n_0+n_1}{n_0}\bar{g}_1 = \frac{n}{n_0}\bar{g}_1$$

Combining the analyses above we find that

$$t = \frac{\bar{g}_1 - \bar{g}_0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} = \frac{n\bar{g}_1/n_0}{\sqrt{\frac{n}{n_0 n_1}\frac{1-r^2}{(n-2)}}} = \frac{\bar{g}_1\sqrt{nn_1/n_0}}{1-r^2}\sqrt{n-2} = \frac{r}{1-r^2}\sqrt{n-2}$$

This is the well known formula for t-statistic for a correlation (under assumption of normality).

ANOVA F-statistic

The ANOVA F-statistic measures the fraction of variation explained by assuming different averages of G across samples where S equal to 0 and 1,

$$F = (n-2)\frac{SSB}{SSW}$$

where

$$SSB = n_0\bar{g}_0^2 + n_1 g_1^2$$

and

$$SSW = \sum_{i=1}^{n_0}(g_i - \bar{g}_0)^2 + \sum_{i=n_0+1}^{n}(g_i - \bar{g}_1)^2$$

Applying the formula derived for t-statistic $\hat{\sigma}^2 = \frac{1-r^2}{n-2}$ we find that

$$SSW = (n-2)\hat{\sigma}^2 = 1 - r^2$$

It follows from $SST = \sum_{i=1}^{n} g_i^2 = 1$ and $SST = SSB + SSW$ that

$$SSB = SST - SSW = r^2$$

Combining the formulas for SSB and SSW yields the equivalence of F and t statistics

$$F = (n-2)\frac{r^2}{1-r^2} = t^2$$

Likelihood ratio test

The likelihood ratio test is defined as doubled difference between data likelihoods under hull and alternative hypotheses

$$LR = 2[l_1(g,s) - l_0(g)]$$

The null model assumes the same normal distribution for samples g_i whether $s_i = 0$ or $s_i = 1$. The likelihood under null model is

$$l_0(g) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\hat{\sigma}_0^2 - \frac{\sum_{i=1}^{n}(g_i - \bar{g})^2}{2\hat{\sigma}_0^2}$$

where

$$\hat{\sigma}_0^2 = \sum_{i=1}^{n} (g_i - \bar{g})^2 / n$$

The alternative model allows the mean of g_i to depend on whether $s_i = 0$ or $s_i = 1$. The likelihood under alternative model is

$$l_1(g, s) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \hat{\sigma}_1^2 - \frac{\sum_{i=1}^{n_0}(g_i - \bar{g}_0)^2 + \sum_{i=n_0+1}^{n}(g_i - \bar{g}_1)^2}{2\hat{\sigma}_1^2}$$

where

$$\hat{\sigma}_1^2 = \left[\sum_{i=1}^{n_0} (g_i - \bar{g}_0)^2 + \sum_{i=n_0+1}^{n} (g_i - \bar{g}_1)^2 \right] / n$$

We can simplify the estimates of variance

$$\hat{\sigma}_0^2 = \sum_{i=1}^{n} (g_i - \bar{g})^2 / n = 1/n$$

$$\hat{\sigma}_1^2 = \left[\sum_{i=1}^{n_0} (g_i - \bar{g}_0)^2 + \sum_{i=n_0+1}^{n} (g_i - \bar{g}_1)^2 \right] / n = SSW/n = (1 - r^2)/n$$

We can simplify the likelihoods using the formulas for $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$

$$l_0(g) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log[1/n] - \frac{n}{2}$$

$$l_1(g, s) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log[(1 - r^2)/n] - \frac{n}{2}$$

This leads to the formula for the likelihood ratio test

$$LR = -\log(1 - r^2)$$

Summary

We can conclude that all four test statistics (r^2, t^2, F, and LR) are monotone functions of each other and thus are equivalent:

$$F = t^2 = \frac{r^2}{1 - r^2}(n - 2) \quad \text{and} \quad LR = -\log(1 - r^2)$$