Exact test for Hardy-Weinberg equilibrium

Exact test: calculate the probability under the null.

H_0 : The genotype frequency follows HWE

Statistical inference is made by asking: what is the cumulative probability of obtaining a sample [with $(2n_{aa} + n_{ab})$ A alleles and $(2n_{bb} + n_{ab})$ B alleles] with a probability at least as low as that of the observed sample under the hypothesis that it was drawn from a population in Hardy-Weinberg equilibrium? This cumulative probability is obtained.

$$P_{HWE} = \sum_{n_{ab^*}} Indicator(P(N_{ab} = n_{ab}) > P(N_{ab} = n^*_{ab}))[P(N_{ab} = n^*_{ab}|n_a, n_b, n)]$$

To be able to solve the above equation, we need,

1. Be able to find all genotype configurations.

2. Be able to calculate $P(N_{ab} = n_{ab})|n_a, n_b, n)$

$$P(N_{ab} = n_{ab})|n_a, n_b, n) = P(n_{ab}, n_a, n_b)/P(n_a)$$

For numerator of the right hand side,

$$P(n_{ab}, n_a, n_b) = \frac{n!}{n_{aa}!n_{ab}!n_{bb}!}(P_a)^{2n_{aa}}(2 * P_a * P_b)^{n_{ab}}(P_b)^{2n_{bb}} = \frac{n!}{n_{aa}!n_{ab}!n_{bb}!}2^{n_{ab}}(P_a)^{2n_{aa}+n_{ab}}(P_b)^{2n_{bb}+n_{ab}} = \frac{n!}{n_{aa}!n_{ab}!n_{bb}!}2^{n_{ab}}(P_a)^{n_a}(P_b)^{n_b}$$

Denominator of right hand side,

$$P(n_a) = \frac{(2n)!}{n_a!n_b!}(P_a)^{n_a}(P_b)^{n_b}$$

Taking the ratio we got $P(N_{ab} = n_{ab})|n_a, n_b, n) = P(n_{ab}, n_a, n_b)/P(n_a) = \frac{n!}{n_{aa}!n_{ab}!n_{bb}!}2^{n_{ab}} * \frac{n_a!n_b!}{(2n)!}$

Relationship between score statistics, Z score and effect size.

Notation

U_j: scaler, U score statistic for association test for variant j. See reference (Danyu Lin, 2011)

V_j: scaler, variance of U_j

v_j: standard deviation of U_j.

Under the $H_0, U_j \sim N(0, V_j)$

$\hat{\beta}_{1j}$: estimate of effect size (the slope) $\beta_1 j$ in simple linear regression for variant j.

Y_i: Phenotype of individual/sample i

X_{ij}: Genotype of individual/sample i at variant j.

Derivations

For convenience we assume that both genotype and phenotype are centered. (Can easy show that this assumption doesn't matter)

From OLS for simple linear regression, where $Y_i \perp Y_j, i \neq j$, and $Var(Y_i) = \sigma^2$ we have

$$\hat{\beta}_{1j} = \frac{Cov(X_j, Y)}{Var(X_j)} = \frac{\sum\limits_i x_{ij} * y_i}{\sum\limits_i x_{ij}^2}$$

$\sum\limits_i x_{ij}^2$ can be viewed as a constant. Thus rewrite $\hat{\beta}_{1j}$ as $\hat{\beta}_{1j} = \sum\limits_i \frac{x_{ij}}{\sum\limits_i x_{ij}^2} y_i$, which is a linear combinations of y_i

Denote $k_i = \frac{x_{ij}}{\sum\limits_i x_{ij}^2}$, then $\hat{\beta}_{1j} = \sum\limits_i k_i y_i$, where k_i can be considered as a constant. Thus the variance of $\hat{\beta}_{1j}$ is

$$Var(\hat{\beta}_{1j}) = Var(\sum_i k_i y_i) = \sum_i Var(k_i y_i) = \sum_i k_i^2 Var(y_i) = \sigma^2 \sum_i k_i^2$$

$$\sum_i k_i^2 = \sum_i (\frac{x_{ij}}{\sum\limits_i x_{ij}^2})^2 = \sum \frac{x_{ij}^2}{(\sum\limits_i x_{ij}^2)^2} = \frac{\sum\limits_i x_{ij}^2}{(\sum\limits_i x_{ij}^2)^2} = \frac{1}{\sum\limits_i x_{ij}^2}$$

Thus $Var(\hat{\beta}_{1j}) = \frac{\sigma^2}{\sum\limits_i x_{ij}^2} = \frac{\sigma^2}{N * Var(X_j)}$ where N is sample size (number of analyzed individuals).

As $\hat{\beta}_{ji}$ is normally distributed (property of OLS estimator), now we can construct the Z statistic/Z score for variant j under the null,

$$Z_j = \frac{\hat{\beta}_{j1}}{(Var(\hat{\beta}_{j1}))^{1/2}} = \frac{\hat{\beta}_{j1}}{(\frac{\sigma^2}{N * Var(X_j)})^{1/2}} = \frac{\hat{\beta}_{j1} * (N * Var(X_j))^{1/2}}{\sigma}$$

Thus we can then get $\hat{\beta}_{j1}$ as

$$\hat{\beta}_{j1} = \frac{Zj}{(N * Var(X_j))^{1/2}} * \sigma$$

Thus if we only consider the first part of the product, $\frac{Zj}{(N * Var(X_j))^{1/2}}$, then our estimate for effect size is in the unit of standard deviation of our phenotype.

Furthermore, we also know that $X_j \sim Bin(2, AF_j)$, where AF_j denotes that allele frequency of the alternative allele at site j. Thus $Var(Xj) = 2 * AF_j * (1 - AF_j)$,

Thus, we now can comprehensively get $\hat{\beta}_{j1}$, namely effect size (in the unit of standard deviation of phenotype) of variant j from summary statistic data (e.g. association results generated by rvtest) as well as the standard deviation of $\hat{\beta}_{j1}$ as follows,

$$\hat{\beta}_{j1} = \frac{Zj}{(N * 2 * AF_j * (1 - AF_j))^{1/2}}$$

$$Var(\hat{\beta}_{j1}) = \frac{1}{N*2*AF_j*(1-AF_j)}$$

where $Z_j = \frac{U_j}{v_j}$.