Power Calculation for Testing If Disease is Associated with Marker in a Case-Control Study Using the GeneticsDesign Package

Weiliang Qiu
email: weiliang.qiu@gmail.com
Ross Lazarus
email: ross.lazarus@channing.harvard.edu

April 16, 2015

1 Introduction

In genetics association studies, investigators are interested in testing the association between disease and DSL (disease susceptibility locus) in a case-control study. However, DSL is usually unknown. Hence what one can do is to test the association between disease and DSL indirectly by testing the association between disease and a marker. The power of the association between disease and DSL depends on the power of the association between disease and the marker, and on the LD (linkage disequilibrium) between disease and the marker.

Assume that both DSL and the marker are biallelic and assume additive model. Given the MAFs (minor allele frequencies) for both DSL and the marker, relative risks, r^2 or D' between DSL and marker, the numbers of cases and controls, and the significant level for hypothesis testing, the tools in the `GeneticsDesign` package can calculate the power for the association test that disease is associated with DSL in a case-control study. This information can help investigator to determine appropriate sample sizes in experiment design stage.

2 Examples

To call the functions in the `R` package GeneticsDesign, we first need to load it into `R`:

```
> library(GeneticsDesign)
```

The following is a sample code to get a table of power for different combinations of high risk allele frequency $Pr(A)$ and genotype relative risk $RR(Aa|aa)$.

```
> set1<-seq(from=0.1, to=0.5, by=0.1)
> set2<-c(1.25, 1.5, 1.75, 2.0)
> len1<-length(set1)
> len2<-length(set2)
> mat<-matrix(0, nrow=len1, ncol=len2)
> rownames(mat)<-paste("MAF=", set1, sep="")
> colnames(mat)<-paste("RRAA=", set2, sep="")
> for(i in 1:len1)
+ { a<-set1[i]
+   for(j in 1:len2)
+   { b<-set2[j]
+     res<-GPC.default(pA=a,pD=0.1,RRAa=(1+b)/2, RRAA=b, Dprime=1,pB=a, quiet=T)
+     mat[i,j]<-res$power
+   }
+ }
> print(round(mat,3))

        RRAA=1.25 RRAA=1.5 RRAA=1.75 RRAA=2
MAF=0.1     0.145    0.400     0.690  0.884
MAF=0.2     0.214    0.594     0.877  0.977
MAF=0.3     0.259    0.682     0.926  0.989
MAF=0.4     0.280    0.711     0.936  0.990
MAF=0.5     0.282    0.702     0.925  0.986
```

A Technical Details

Suppose that A is the minor allele for the disease locus; a is the common allele for the disease locus; B is the minor allele for the marker locus; and b is the common allele for the marker locus. We use D to denote the disease status "diseased" and use $\bar{D}$ to denote the disease status "not diseased".

Given the minor allele frequencies $Pr(A)$ and $Pr(B)$ and Linkage disequilibrium (LD) measure $\mathbf{D}'$, we can get haplotype frequencies $Pr(AB)$, $Pr(Ab)$, $Pr(aB)$, and $Pr(ab)$:

$$\begin{aligned} Pr(AB) &= Pr(A)Pr(B) + \mathbf{D} \\ Pr(aB) &= Pr(a)Pr(B) - \mathbf{D} \\ Pr(Ab) &= Pr(A)Pr(b) - \mathbf{D} \\ Pr(ab) &= Pr(a)Pr(b) + \mathbf{D}, \end{aligned}$$

where

$$\mathbf{D} = \mathbf{D}' d_{\max}$$

and

$$d_{\max} = \begin{cases} \min\left[Pr(A)Pr(b), Pr(a)Pr(B)\right], & \text{if } \mathbf{D} > 0, \\ \max\left[-Pr(A)Pr(B), -Pr(a)Pr(b)\right], & \text{if } \mathbf{D} < 0. \end{cases}$$

That is,

$$\mathbf{D} = Pr(AB) - Pr(A)Pr(B).$$

Note that $\mathbf{D} > 0$ means $Pr(AB) > Pr(A)Pr(B)$, i.e., the probability of occuring the haplotype AB is higher than the probability that the haplotype AB occurs merely by chance. $\mathbf{D} < 0$ means $Pr(AB) < Pr(A)Pr(B)$, i.e., the probability of occuring the haplotype AB is smaller than the probability that the haplotype AB occurs merely by chance. In other words, given the same sample size, we can observe haplotype AB more often when $\mathbf{D} > 0$ than when $\mathbf{D} < 0$. Hence the power of association test that marker is associated with disease is larger when $\mathbf{D} > 0$ than when $\mathbf{D} < 0$. **Hence, we assume that $\mathbf{D} > 0$.**

$\mathbf{D}$ can also be rewritten as

$$\mathbf{D} = Pr(B|A)Pr(A) - Pr(A)Pr(B) = \left[Pr(B|A) - Pr(B)\right] Pr(A).$$

Hence $\mathbf{D} > 0$ is equivalent to $Pr(B|A) > Pr(A)$ which indicates positive association between minor allele A in disease locus and minor allele B in marker locus.

The relative risks are

$$RR(AA|aa) = \frac{Pr(D|AA)}{Pr(D|aa)}$$
$$RR(Aa|aa) = \frac{Pr(D|Aa)}{Pr(D|aa)}$$

The disease prevalence $Pr(D)$ can be rewritten as

$$Pr(D) = Pr(D|AA)Pr(AA) + Pr(D|Aa)Pr(Aa) + Pr(D|aa)Pr(aa).$$

Dividing both sides by $Pr(D|aa)$, we get

$$\begin{aligned} \frac{Pr(D)}{Pr(D|aa)} &= \frac{Pr(D|AA)}{Pr(D|aa)}Pr(AA) + \frac{Pr(D|Aa)}{Pr(D|aa)}Pr(Aa) + \frac{Pr(D|aa)}{Pr(D|aa)}Pr(aa) \\ &= RR(AA|aa)\left[Pr(A)\right]^2 + RR(Aa|aa)2Pr(A)Pr(a) + \left[Pr(a)\right]^2. \end{aligned}$$

Hence

$$\begin{aligned} Pr(D|aa) &= \frac{Pr(D)}{RR(AA|aa)\left[Pr(A)\right]^2 + RR(Aa|aa)2Pr(A)Pr(a) + \left[Pr(a)\right]^2} \\ Pr(D|Aa) &= RR(Aa|aa)Pr(D|aa) \\ Pr(D|AA) &= RR(AA|aa)Pr(D|aa) \end{aligned}$$

Once we obtain the penetrances of disease locus ($Pr(D|aa)$, $Pr(D|Aa)$, and $Pr(D|AA)$), we can get the sampling probabilities:

$$\begin{aligned} Pr(BB|D) &= \frac{Pr(D, BB)}{Pr(D)} \\ &= \frac{Pr(D, BB, AA) + Pr(D, BB, Aa) + Pr(D, BB, aa)}{Pr(D)} \\ &= \frac{Pr(D|BB, AA)Pr(BB, AA) + Pr(D|BB, Aa)Pr(BB, Aa) + Pr(D|BB, aa)Pr(BB, aa)}{Pr(D)} \end{aligned}$$

$$\begin{aligned} Pr(Bb|D) &= \frac{Pr(D, Bb)}{Pr(D)} \\ &= \frac{Pr(D, Bb, AA) + Pr(D, Bb, Aa) + Pr(D, Bb, aa)}{Pr(D)} \\ &= \frac{Pr(D|Bb, AA)Pr(Bb, AA) + Pr(D|Bb, Aa)Pr(Bb, Aa) + Pr(D|Bb, aa)Pr(Bb, aa)}{Pr(D)} \end{aligned}$$

$$\begin{aligned} Pr(bb|D) &= \frac{Pr(D, bb)}{Pr(D)} \\ &= \frac{Pr(D, bb, AA) + Pr(D, bb, Aa) + Pr(D, bb, aa)}{Pr(D)} \\ &= \frac{Pr(D|bb, AA)Pr(bb, AA) + Pr(D|bb, Aa)Pr(bb, Aa) + Pr(D|bb, aa)Pr(bb, aa)}{Pr(D)} \end{aligned}$$

We assume that

$$Pr(D|BB, AA) = Pr(D|AA), \quad Pr(D|BB, Aa) = Pr(D|Aa), \quad Pr(D|BB, aa) = Pr(D|aa).$$

We also have

$$\begin{aligned} Pr(BB, AA) &= [Pr(AB)]^2 \\ Pr(BB, Aa) &= 2Pr(AB)Pr(aB) \\ Pr(BB, aa) &= [Pr(aB)]^2 \end{aligned}$$

$$\begin{aligned} Pr(Bb, AA) &= 2Pr(AB)Pr(Ab) \\ Pr(Bb, Aa) &= 2\left[Pr(AB) * Pr(ab) + Pr(Ab)Pr(aB)\right] \\ Pr(Bb, aa) &= 2Pr(aB)Pr(ab) \end{aligned}$$

$$\begin{aligned} Pr(bb, AA) &= [Pr(Ab)]^2 \\ Pr(bb, Aa) &= 2Pr(Ab)Pr(ab) \\ Pr(bb, aa) &= [Pr(ab)]^2 \end{aligned}$$

Hence

$$
\begin{aligned}
Pr(BB|D) &= \frac{Pr(D|BB,AA)Pr(BB,AA) + Pr(D|BB,Aa)Pr(BB,Aa) + Pr(D|BB,aa)Pr(BB,aa)}{Pr(D)} \\
&= \frac{Pr(D|AA)\left[Pr(AB)\right]^2 + Pr(D|Aa)\left[2Pr(AB)Pr(aB)\right] + Pr(D|aa)\left[Pr(aB)\right]^2}{Pr(D)} \\
Pr(Bb|D) &= \frac{Pr(D|Bb,AA)Pr(Bb,AA) + Pr(D|Bb,Aa)Pr(Bb,Aa) + Pr(D|Bb,aa)Pr(Bb,aa)}{Pr(D)} \\
&= \frac{Pr(D|AA)\left[2Pr(AB)Pr(Ab)\right] + Pr(D|Aa)\left\{2\left[Pr(AB)Pr(ab) + Pr(Ab)Pr(aB)\right]\right\}}{Pr(D)} \\
&+ \frac{Pr(D|aa)\left[2Pr(aB)Pr(ab)\right]}{Pr(D)} \\
Pr(bb|D) &= \frac{Pr(D|bb,AA)Pr(bb,AA) + Pr(D|bb,Aa)Pr(bb,Aa) + Pr(D|bb,aa)Pr(bb,aa)}{Pr(D)} \\
&= \frac{Pr(D|AA)\left[Pr(Ab)\right]^2 + Pr(D|Aa)\left[2Pr(Ab)Pr(ab)\right] + Pr(D|aa)\left[Pr(ab)\right]^2}{Pr(D)}
\end{aligned}
$$

To calculate the sampling probabilities $Pr(BB|\bar{D})$, $Pr(Bb|\bar{D})$, and $Pr(bb|\bar{D})$, we can apply Bayesian rules again:

$$
\begin{aligned}
Pr(BB|\bar{D}) &= \frac{Pr(\bar{D}|BB)Pr(BB)}{Pr(\bar{D})} = \frac{[1 - Pr(D|BB)][Pr(B)]^2}{1 - Pr(D)} \\
Pr(Bb|\bar{D}) &= \frac{Pr(\bar{D}|Bb)Pr(Bb)}{Pr(\bar{D})} = \frac{[1 - Pr(D|Bb)]2Pr(B)Pr(b)}{1 - Pr(D)} \\
Pr(bb|\bar{D}) &= \frac{Pr(\bar{D}|bb)Pr(bb)}{Pr(\bar{D})} = \frac{[1 - Pr(D|bb)][Pr(b)]^2}{1 - Pr(D)}
\end{aligned}
$$

and

$$
\begin{aligned}
Pr(D|BB) &= \frac{Pr(BB|D)Pr(D)}{Pr(BB)} = \frac{Pr(BB|D)Pr(D)}{\left[Pr(B)\right]^2} \\
Pr(D|Bb) &= \frac{Pr(Bb|D)Pr(D)}{Pr(Bb)} = \frac{Pr(Bb|D)Pr(D)}{[2Pr(B)Pr(b)]} \\
Pr(D|bb) &= \frac{Pr(bb|D)Pr(D)}{Pr(bb)} = \frac{Pr(bb|D)Pr(D)}{\left[Pr(b)\right]^2}.
\end{aligned}
$$

The expected allele frequencies are

$$\begin{aligned}
Pr(B|D) &= \frac{2n_{case}Pr(BB|D) + n_{case}Pr(Bb|D)}{2n_{case}} \\
&= Pr(BB|D) + Pr(Bb|D)/2 \\
Pr(b|D) &= \frac{2n_{case}Pr(bb|D) + n_{case}Pr(Bb|D)}{2n_{case}} \\
&= Pr(bb|D) + Pr(Bb|D)/2 \\
Pr(B|\bar{D}) &= Pr(BB|\bar{D}) + Pr(Bb|\bar{D})/2 \\
Pr(b|\bar{D}) &= Pr(bb|\bar{D}) + Pr(Bb|\bar{D})/2.
\end{aligned}$$

The counts for cases

$$n_{2c} = n_{cases}Pr(BB|D), \quad n_{1c} = n_{cases}Pr(Bb|D), \quad n_{0c} = n_{cases}Pr(bb|D).$$

The counts for controls

$$n_{2n} = n_{controls}Pr(BB|\bar{D}), \quad n_{1n} = n_{controls}Pr(Bb|\bar{D}), \quad n_{0n} = n_{controls}Pr(bb|\bar{D}).$$

Table 1: Counts for cases and controls

marker genotype	cases	controls
BB	n_{2c}	n_{2n}
Bb	n_{1c}	n_{1n}
bb	n_{0c}	n_{0n}
total	n_{cases}	$n_{controls}$

Odds ratios are

$$\begin{aligned}
OR(BB|bb) &= \frac{n_{2c}n_{0n}}{n_{2n}n_{0c}} \\
&= \frac{Pr(BB|D)Pr(bb|\bar{D})}{Pr(BB|\bar{D})Pr(bb|D)} \\
OR(Bb|bb) &= \frac{n_{1c}n_{0n}}{n_{1n}n_{0c}} \\
&= \frac{Pr(Bb|D)Pr(bb|\bar{D})}{Pr(Bb|\bar{D})Pr(bb|D)}
\end{aligned}$$

Define

$$U = \sum_{i=1}^{3} x_i \left(\frac{S}{N} r_i - \frac{R}{N} s_i \right),$$

where

$$\begin{aligned} R =& n_{0c} + n_{1c} + n_{2c}, S = n_{0n} + n_{1n} + n_{2n}, N = R + S \\ .x_1 =& 2,\ x_2 = 1,\ x_3 = 0, \\ r_1 =& n_{0c},\ r_2 = n_{1c},\ r_3 = n_{2c}, \\ s_1 =& n_{0n},\ s_2 = n_{1n},\ s_3 = n_{2n}. \end{aligned}$$

We have

$$V = \widehat{Var}(U) = \frac{RS}{N^3}\left[N\sum_{i=1}^{3} x_i^2 n_i - \left(\sum_{i=1}^{3} x_i n_i\right)^2\right].$$

The Linear Trend Test Statistic[1]

$$Z_T^2 = \left(\frac{U}{\sqrt{V}}\right)^2$$

asymptotically follows a χ_1^2 distribution under the null hypothesis. Under alternative hypothesis, Z_T^2 asymptotically follows a non-central chi-square distribution χ_{1,λ_T}^2 with non-centrality parameter

$$\lambda_T = RS\frac{\left[\sum_{i=1}^{3} x_i(p_{1i} - p_{0i})\right]^2}{\sum_{i=1}^{3} x_i^2(Rp_{0i} + Sp_{1i}) - \left[\sum_{i=1}^{3} x_i(Rp_{0i} + Sp_{1i}\right]^2/N},$$

where

$$\begin{aligned} x_1 =& 2,\ x_2 = 1,\ x_3 = 0, \\ n_1 =& n_{2c} + n_{2n},\ n_2 = n_{1c} + n_{1n},\ n_3 = n_{0c} + n_{0n}, \\ p_{01} =& Pr(BB|D),\ p_{02} = Pr(Bb|D),\ p_{03} = Pr(bb|D), \\ p_{11} =& Pr(BB|\bar{D}),\ p_{12} = Pr(Bb|\bar{D}),\ p_{13} = Pr(bb|\bar{D}). \end{aligned}$$

Given significant level α, the power $1-\beta$ satisfies:

$$\begin{aligned} Pr(Z_T^2 > c|H_0) =& \alpha \\ Pr(Z_T^2 > c|H_a) =& 1-\beta. \end{aligned}$$

References

[1] Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375-386, 1955.

[2] Cochran, W. G. Some methods for strengthening the common chi-squared tests. *Biometrics*, 10, 417-451, 1954.

[1] http://linkage.rockefeller.edu/pawe3d/help/Linear-trend-test-ncp.html

[3] Gordon D. and Finch S. J. and Nothnagel M. and Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum. Hered.*, 54:22-33, 2002.

[4] Gordon D. and Haynes C. and Blumenfeld J. and Finch S. J. PAWE-3D: visualizing Power for Association With Error in case/control genetic studies of complex traits. *Bioinformatics*, 21:3935-3937, 2005.

[5] Purcell S. and Cherny S. S. and Sham P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex trait *Bioinformatics*, 19(1):149-150, 2003.

[6] Sham P. *Statistics in Human Genetics.* Arnold Applications of Statistics, 1998.