

Effective Sample Size of Affected Sib Pairs in Case-Control Analyses

*Wentian Li¹, Ph.D, Yaning Yang², Ph.D, Peter K. Gregersen¹,
M.D.*

1. Robert S Boas Center for Genomics & Human Genetics
North Shore LIJ Institute for Medical Research
2. Department of Statistics and Finance
University of Science and Technology of China

sept 2005 nslij institute retreat

Abstract

Many sibpair collections have been established for linkage mapping studies. These collections can be valuable resources for association studies, but the correlation among siblings must be appropriately allowed for in the analysis to obtain valid tests. We propose an “effective sample size” method to capture the correlation among genotypes of sibs from the same pedigree. This method is based on a theoretical calculation of standard errors and variances of the genotype frequency when siblings are included in the dataset. When the effective sample size method is applied to case-control analysis, where control samples are independent but case samples are collected from the affected-sib-pairs, the confidence interval of odd-ratio can be approximatedly calculated. This calculation is straightforward to carry out and thus simplifies the procedure for including both sibs, from previously collected affected-sib-pair pedigrees, in a case-control analysis. The method extends in studies with mixed sibpairs and singletons for cases and controls. The method is implemented for the scientific community at an online web site.

Motivation

Genetic association analysis is applied to a group of independently collected patients and normal persons; whereas genetic linkage analysis is applied to pedigree data with interconnected individuals.

If the two types of analyses are carried out as two separate projects, no problem. However, we want to use the pedigree data collected for linkage analysis for association analysis also, to save the cost of collecting new patients.

Example

Most of the North American Rheumatoid Arthritis Consortium (NARAC) data consists of pedigrees with two siblings with Rheumatoid Arthritis. These are “affected sib pairs” (ASPs).

Clearly, the two sibs are genetically correlated, and using both of them violates the assumption of “independent samples” in statistical tests. One solution (the old one) is to randomly pick one sib from a sibpair. This is not efficient as many samples are discarded.

Main Questions

- = Can we use both affected sibs as case samples in a case-control association analysis?
- = If we do, what is the correct procedure for a statistical test?

Quick Answers

- = Yes, we can use both affected sibs.
- = Instead of two samples, the “effective sample size” of two sibs is roughly 1.4. Using this 1.4 for subsequent calculations.

An important observation: **Correlation among samples does not change (bias) the calculation of mean (average), it only affects the calculation of variance (standard deviation).**

Two quantities commonly calculated in a case-control analysis, (1) 95% confidence interval of the odd-ratio, and (2) p -value of the Pearson's chi-square test, rely on the variance. Consequently, these two calculations should be affected by the correlation among samples.

Review Basic Case-Control Analysis: (1) Odd Ratio and Its Confidence Interval

	<i>allele-a</i>	<i>allele-A</i>
<i>case</i>	N_{11}	N_{12}
<i>control</i>	N_{21}	N_{22}

Odd Ratio (OR): $OR = \frac{N_{11}N_{22}}{N_{12}N_{21}}$

95% Confidence Interval (CI) of odd ratio:

from $\exp(\log(OR) - 1.96\sigma)$ to $\exp(\log(OR) + 1.96\sigma)$, where

$$\sigma = \sqrt{1/N_{11} + 1/N_{12} + 1/N_{21} + 1/N_{22}}$$

Review of Basic Case-Control Analysis: (2) Chi-square Tests

X^2 test statistic:

$$X^2 = \frac{(N_{11}N_{22} - N_{12}N_{21})^2(N_{11} + N_{12} + N_{21} + N_{22})}{(N_{11} + N_{12})(N_{21} + N_{22})(N_{11} + N_{21})(N_{12} + N_{22})}$$

p -value of chi-square test: plugging X^2 into the χ^2 distribution with 1 degree of freedom; the tail area is the p -value.

A statistic test is significant when p -value is smaller than certain threshold (e.g. 0.05), or, if the whole CI of OR is larger than 1.

The Effective Sample Size Method in Handling Affected Sibpairs

Replacing N_{11} and N_{12} with $N_{11,eff} = \alpha N_{11}$ and $N_{12,eff} = \alpha N_{12}$, where $\alpha \approx 0.7096$.

Then plugging $N_{11,eff}$ and $N_{12,eff}$ into both 95%CI of OR and X^2 formula, derive new the confidence interval and the p -value.

For example, if $N_{11} = 362$, $N_{12} = 83$, $N_{21} = 434$, $N_{22} = 66$, using the apparent sample size leads to 95% CI= (0.466, 0.943) and p -value= 0.022 (significant at the 0.05 level). Using the effective sample size $N_{11,eff} = 256.8752$ and $N_{12,eff} = 58.8968$ leads to 95% CI= (0.342, 0.973) and p -value= 0.0352. The change from $p=0.022$ to $p=0.035$ becomes important only if the threshold for significance is set at 0.03.

Mathematical Justification of the Effective Sample Size Method

Variance of genotype frequencies for sibpair samples can be calculated analytically, for example, for the homozygous genotype aa (p is the frequency of allele a): $Var_{aa} = \frac{p^2 - 2p^4 + (1+p)^2 p^2 / 4}{2N_{ped}}$ *

which should be equal to $Var_{aa} = \frac{p^2(1-p^2)}{N_{eff}}$

In other words, the concept of “effective sample size” is the sample size used in the formula *as if the samples are independent*.

*Details of the derivation can be found in the paper.

Mathematical Justification of the Effective Sample Size Method (cont.)

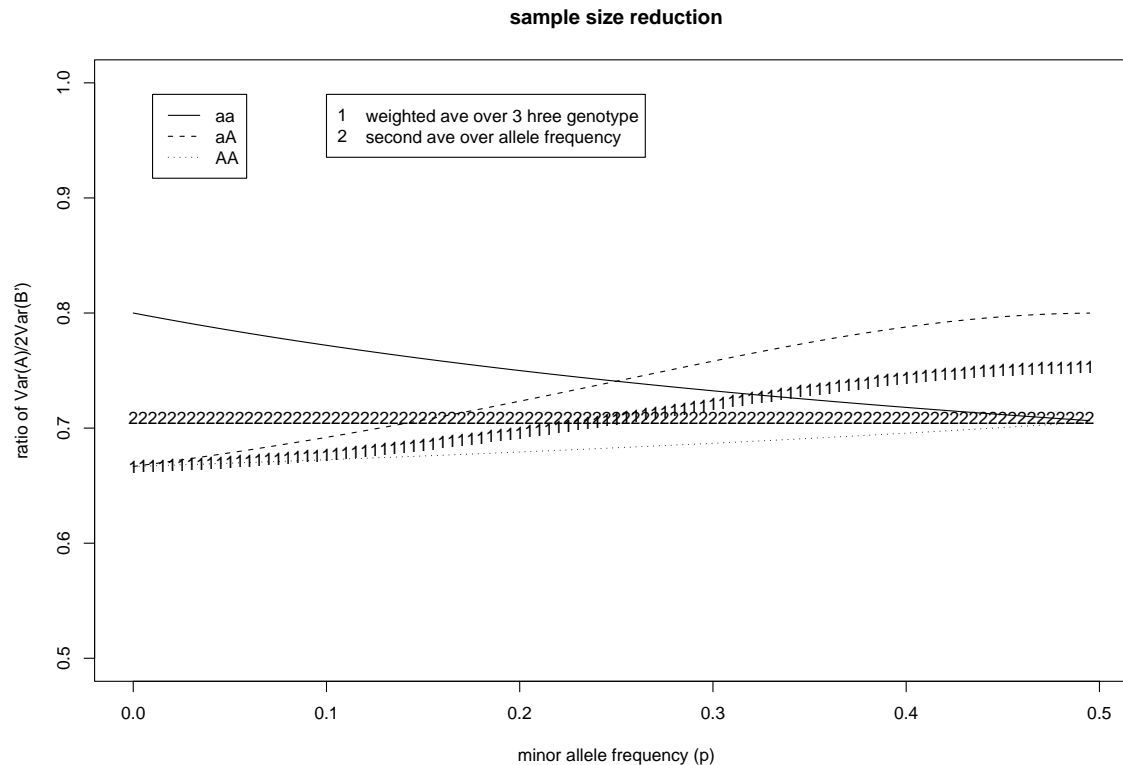
However, setting the two Var to be equal to derive

$$\frac{N_{eff}}{2N_{ped}} = \frac{p^2 - 2p^4 + (1+p)^2 p^2 / 4}{p^2 - p^4} = \frac{1 - 2p^2 + (1+p)^2 / 4}{1 - p^2}$$

encountered two problems:

- (1) The similar equations for genotypes aA and AA lead to different values of N_{eff} .
- (2) The value of N_{eff} depends on the allele frequency p .

The following plot provides the key in answering these two questions.



Sample size reduction ($\alpha = N_{\text{eff}}/(2N)$ for scheme-B') as estimated by the ratio of two variances: the variance in scheme-A (divided by 2), and the variance in scheme-B'. The line labeled by "1" represents the weighted average of the α for the three genotypes ($\bar{\alpha}(p)$). The horizontal line labeled by "2" represents the second round of average of $\bar{\alpha}(p)$ over all allele frequencies ($\bar{\bar{\alpha}}$).

Mathematical Justification of the Effective Sample Size Method (cont.)

The previous plot shows that the influence of allele frequency on variance ratio (thus effective sample size over actual sample size ratio) is limited; also the three genotypes show similar (though not identical) curves.

As an approximation, we can use the double average (first, averaging over three genotypes, second, averaging over the allele frequency) to obtain the effective sample size, this becomes $N_{eff} = 0.7096N_{actual}$.

Conclusions

- = Both affected sibs can be used as case samples in a case-control analysis
- = A simple treatment of the correlation between sibs is to use the effective sample of 0.7096 for each sib. Then recalculate both 95% confidence interval of odd-ratio and chi-square test statistic (and p -value).
- = Typically, both 95% CIs of OR and p -values are only moderately affected due to the correlation between two sibs.

Reference

W Li, Y Yang, PK Gregersen (2005), “Effective sample size of affected sib pairs in case-control analyses”, submitted.

Web Resources

ESSCC (Effective Sample Size calculator for Case-Control analysis) <http://www.nslj-genetics.org/esscc/>

A bibliography on linkage disequilibrium analysis:

<http://www.nslj-genetics.org/ld/>

A comprehensive list of genetic analysis programs:

<http://www.nslj-genetics.org/soft/>