Lecture 1: Case-Control Association Testing

Instructors: Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2021

Introduction

- Association mapping is now routinely being used to identify loci that are involved with complex traits.
- Technological advances have made it feasible to perform case-control association studies on a genome-wide basis with hundreds of thousands of markers in a single study.
- We consider testing a genetic marker for association with a disease in a sample of unrelated subjects.
- Case-control association methods essentially test for independence between trait and allele/genotype.

Case-Control Association Testing

- Allelic Association Tests
  - Allele is treated as the sampling unit
  - Typically make an assumption of Hardy-Weinberg equilibrium (HWE). Alleles within an individual are conditionally independent, given the trait value.
- Genotypic Association Tests
  - Individual is the sampling unit
  - Does not assume HWE

Pearson's χ^2 Test for Allelic Association

- The classical Pearson's χ^2 test is often used for allelic association testing.
- This test looks for deviations from independence between the trait and allele.
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled “1” and “2”
- Let N_{ca} be the number of cases and N_{co} be the number of controls with genotype data at the marker.

Pearson's χ^2 Test for Allelic Association

- Below is a 2×2 contingency table for trait and allelic type

	Cases	Controls	Total
Allele 1	n_1^{ca}	n_1^{co}	n_1
Allele 2	n_2^{ca}	n_2^{co}	n_2
Total	$2N_{ca}$	$2N_{co}$	T

- n_1^{ca} is the number of type 1 alleles in the cases and $n_1^{ca} = 2 \times$ the number of homozygous $(1, 1)$ cases $+$ the number of heterozygous (1,2) cases
- n_2^{co} is the number of type 2 alleles in the controls and $n_2^{co} = 2$ $\times$ the number of homozygous $(2, 2)$ controls $+$ the number of heterozygous (1,2) controls
- Hypotheses
  - H_0: there is *no association* between the row variable and column variable
  - H_a: there *is* an association between the two variables

Pearson's χ^2 Test for Allelic Association

- Can use Pearson's χ^2 test for independence. The statistic is:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed cell} - \text{ Expected cell})^2}{\text{Expected cell}}$$

- What is the the expected cell number under H_0? For each cell, we have

$$\text{Expected Cell Count} = \frac{\text{row total } \times \text{ col total}}{\text{total count}}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1) = (2-1)(2-1) = 1$ degree of freedom

LHON Example: Pearson's χ^2 Test

- From Phasukijwattana et al. (2010), Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- Corresponding 2×2 contingency table for allelic type and case-control status

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Intuition for the test: Suppose H_0 is true, allelic type and case-control status are independent, then what counts would we expect to observe?
- Recall that under the independence assumption

LHON Example: Pearson's χ^2 Test

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Let n be the total number of alleles in the study. Assuming independence, the expected number of case alleles that are of type T is:

$$n \times P(\text{Allele is from a Case and Allelic type is T})$$

$$= nP(\text{Allele is from a Case})P(\text{Allelic type is T})$$

$$= 656\left(\frac{178}{656}\right)\left(\frac{550}{656}\right) = \frac{(178)(550)}{656} = 149.2378$$

LHON Example: Pearson's χ^2 Test

- Fill in the remaining cells for the expected counts

	Cases	Controls	Total
Allele T	149.2378		
Allele C			
Total			

- Calculate the X^2 statistic

$$X^2 = \frac{(158 - 149.2378)^2}{149.2378} + \cdots + \frac{(86 - 77.2378)^2}{77.2378} = 4.369$$

- What is the p-value?

$$P(\chi_1^2 \geq 4.369) = .037$$

LHON Example: Pearson's χ^2 Test

- Fill in the remaining cells for the expected counts

	Cases	Controls	Total
Allele T	149.2378	400.7622	550
Allele C	28.7622	77.2378	106
Total	178	478	656

- Calculate the X^2 statistic

$$X^2 = \frac{(158 - 149.2378)^2}{149.2378} + \cdots + \frac{(86 - 77.2378)^2}{77.2378} = 4.369$$

- What is the p-value?

$$P(\chi_1^2 \geq 4.369) = .037$$

Fisher's Exact Test for Allelic Association

- For contingency tables that have cells with less than 5 observations
- Consider the table below

	Cases	Controls	Total
Allele T	21	14	35
Allele C	3	10	13
Total	24	24	48

- The marginal counts of the table are fixed: There are 24 case alleles, 24 control alleles, 35 T alleles, and 13 C alleles
- Let X be the number of case alleles that are of type T. A test based on X can be constructed.
- Under the null hypothesis, X will have a hypergeometric distribution where the probability that $X = x$ is

$$\binom{35}{x}\binom{13}{24-x}\Big/\binom{48}{24}$$

Fisher's Exact Test for Allelic Association

- Obtain the probability distribution for X

x	11	12	13	14	15	16	17	18	19	20	21	22	23	24
P(X=x)														

- $P(X = 11)$ is

$$\binom{35}{11}\binom{13}{13}\Big/\binom{48}{24} = .00001$$

- $P(X = 12)$ is

$$\binom{35}{12}\binom{13}{12}\Big/\binom{48}{24} = .0003$$

Fisher's Exact Test for Allelic Association

- Obtain the probability distribution for X

x	P(X=x)
11	.00001
12	.0003
13	.004
14	.021
15	.072
16	.162
17	.241
18	.241
19	.162
20	.072
21	.021
22	.004
23	.0003
24	.00001

- Construct a rejection region for a two-sided test with $\alpha = .05$.
- Can the null hypothesis be rejected at the .05 level for the observed value $X = 21$?

Fisher's Exact Test for Allelic Association

- Obtain the probability distribution for X

x	P(X=x)
11	.00001
12	.0003
13	.004
14	.021
15	.072
16	.162
17	.241
18	.241
19	.162
20	.072
21	.021
22	.004
23	.0003
24	.00001

- So, a rejection region for a two-sided test with $\alpha = .05$ would consist of the following values for X: 11, 12, 13, 14, 21, 22, 23, and 24.
- The observed X value of 21 for the data falls in this region, so the test would reject at the level .05.

The Armitage Trend Test for Genotypic Association

- The most common genotypic test for unrelated individuals is the Armitage trend test (Sasieni 1997)
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled "1" and "2"
- Let $Y_i = 2$ if individual i is homozygous (1,1), 1 if the i is heterozygous, and 0 if i is homozygous (2,2)
- Let $X_i = 1$ if i is a case and 0 if i is a control.
- A simple linear regression model of

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

The Armitage Trend for Genotypic Association

- To test this hypothesis, the Armitage trend test statistic is

$$A_r = \frac{\hat{\beta}_1^2}{VAR(\hat{\beta}_1)} = Nr_{xy}^2$$

  where r_{xy}^2 is the squared correlation between genotype variable Y and phenotype variable X.
- Note that the variance estimate for Y that is used in the calculation of the Armitage trend test is the sum of the squared deviations of Y from the fitted values of Y for regression with only an intercept term.
- Under the null hypothesis, A_r will follow an approximate χ^2 distribution with 1 degree of freedom.
- The Armitage trend test can be shown to be valid when HWE does not hold.

LHON Example: Armitage Trend Test

- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- The Armitage test statistic for this data is

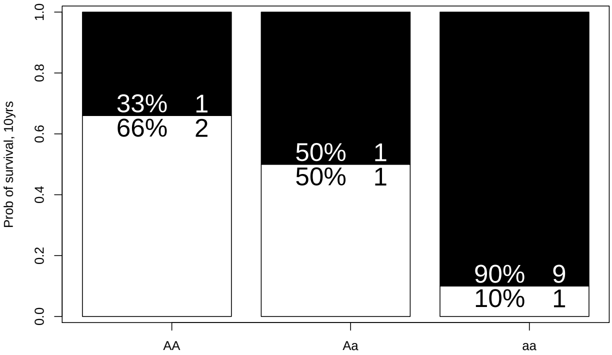
$$A_r = Nr_{xy}^2 = 328(.0114) = 3.74$$

- The *p*-value is

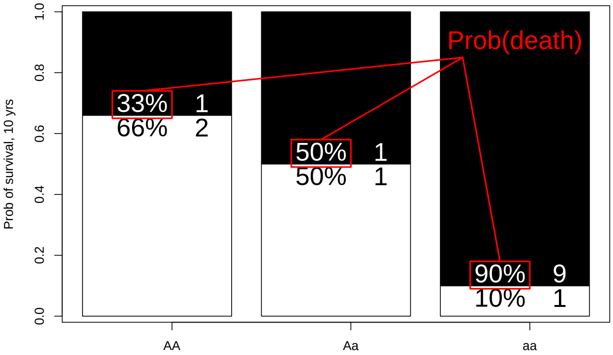
$$P(\chi_1^2 \geq 3.743) = .053$$

Odds Ratios (ORs) for Genotypes

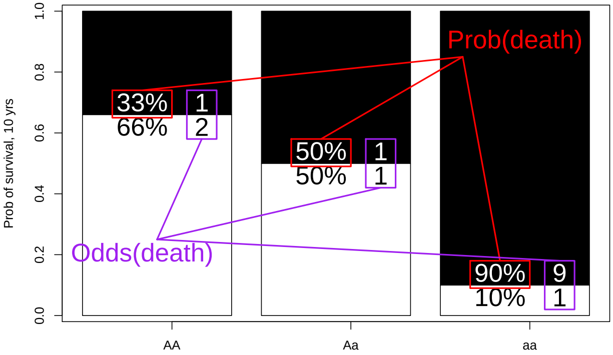
- What are **odds**? Really just **probability**...
- Odds are a [gambling-friendly] measure of chance;



Odds Ratios (ORs) for Genotypes



Odds Ratios (ORs) for Genotypes



– so what are **odds ratios**?

Odds Ratios (ORs) for Genotypes

	Cases	Controls
TT	A	B
CT	A'	B'
CC	C	D

- Typically choose a reference genotype.

$$OR_{TT} = \frac{\text{odds of disease in an individual with the TT genotype}}{\text{odds of disease in an individual with the CC genotype}}$$

$$OR_{CT} = \frac{\text{odds of disease in an individual with the CT genotype}}{\text{odds of disease in an individual with the CC genotype}}$$

- $OR_{TT} = 1$ implies no association with and disease. Similarly for OR_{CT}.
- $OR_{TT} > 1$ or $OR_{TT} < 1$ implies association with the disease.

Odds Ratios (ORs) for Genotypes

- Logistic regression is generally used to get odds ratios and confidence intervals for genotypes.
- Let π_i be the probability that individual i is affected with the disease and let G_i be the genotype for individual i at the SNP:

$$\text{log(odds of disease for individual } i|G_i)$$

$$= log\left(\frac{\pi_i}{1-\pi_i}\Big|G_i\right)$$

$$= \beta_0 + \beta_{CT} I\{G_i = CT\} + \beta_{TT} I\{G_i = TT\}$$

where $I\{G_i = CT\}$ is 1 if $G_i = CT$ and 0 otherwise, and similarly for $I\{G_i = TT\}$.

Odds Ratios (ORs) for Genotypes

- The coefficient estimates for $\hat{\beta}_{CT}$ and $\hat{\beta}_{TT}$ can be used to calculate odds ratios:

$$OR_{CT} = exp(\hat{\beta}_{CT})$$

$$OR_{TT} = exp(\hat{\beta}_{TT})$$

- 95% CI for OR_{CT} is

$$exp(\hat{\beta}_{CT} \pm 1.96 \times s.e.(\hat{\beta}_{CT}))$$

Odds Ratios for LHON Example

- Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- We will use the R software package to obtain odds ratios and confidence intervals for this data set (as well as Pearson's χ^2 test and Armitage Trend tests).
- Exercises and some commands for analyzing the LHON data with R can be found on the following webpage:

  http://faculty.washington.edu/tathornt/SISG_MODULE8.html

Odds Ratios (ORs) based on Allele Counting

- We can also obtain allelic odds ratios
- Odds ratios based on an allele counting model essentially assumes an additive model
- Genotype TT has twice the risk (or protection) of heterozygous genotype CT.
- Same risk (or protection) for the comparison of heterozygous CT genotype and homozygous CC genotype.

	Cases	Controls
T	n_A	n_B
C	n_C	n_D

$$OR_T = \frac{\text{odds of disease with T allele}}{\text{odds of disease with C allele}}$$

$$= \frac{(n_A/n_B)}{(n_C/n_D)} = \frac{n_A \times n_D}{n_B \times n_C}$$

Odds Ratios (ORs) Allele Counting

	Cases	Controls
T	n_A	n_B
C	n_C	n_D

- $OR_T = 1$ implies no association with and disease
- $OR_T > 1$ or $OR_T < 1$ implies association with the disease

Confidence Intervals for Odds Ratios (ORs)

	Cases	Controls
T	n_A	n_B
C	n_C	n_D

$$OR = \frac{n_A \times n_D}{n_B \times n_C}$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{n_C} + \frac{1}{n_D}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

- Upper limit of 95% CI

$$= exp(log(OR) + 1.96 \times s.e.(log(OR)))$$

Confidence Intervals for Odds Ratios (ORs)

rs6767450	Cases	Controls
T	158	392
C	20	86

$$OR = \frac{n_A \times n_D}{n_B \times n_C}$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{n_C} + \frac{1}{n_D}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

- Upper limit of 95% CI

$$= exp(log(OR) + 1.96 \times s.e.(log(OR)))$$

LHON Example: Confidence Intervals for Odds Ratios (ORs)

rs6767450	Cases	Controls
T	158	392
C	20	86

$$OR = \frac{158 \times 86}{392 \times 20} = 1.7332$$

$$s.e.(log(OR)) = \sqrt{\frac{1}{158} + \frac{1}{392} + \frac{1}{20} + \frac{1}{86}}$$

- Lower limit of 95% CI

$$= exp(log(OR) - 1.96 \times s.e.(log(OR)))$$

$$= exp(log(1.7332) - 1.96 \times 0.2665) = 1.03$$

- Upper limit of 95% CI $= 2.92$

References

- Phasukijwattana N, Kunhapan B, Stankovich J, Chuenkongkaew WL, Thomson R, Thornton T, Bahlo M, Mushiroda T, Nakamura Y, Mahasirimongkol S, et al. (2010). Genome-wide linkage scan and association study of PARL to the expression of LHON families in Thailand. *Hum. Genet.* **128**, 39-49.
- Sasieni P (1997). From genotypes to genes: doubling the sample size. *Biometrics* **5**, 1254-1261.