6. Genetic association analysis

Iván Galván-Femenía [1], Jan Graffelman [1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

December 15, 2016

ivan.galvan@upc.edu

Master in Innovation and Research in Informatics (MIRI)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Contents

Genetic association studies

Goal:

- Investigate associations between markers and a trait (disease).

Designs:

- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

We will focus on population-based association studies

Preliminaries

- The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous).
- $Y_i = 1$ individual i has the trait, $Y_i = 0$, individual i does not have the trait.
- The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

The data table

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

We can test for association using different genetic models:

- A codominant model
- A dominant model
- A recessive model
- An additive model

Codominant test

- We test the null hypothesis of no effect of the marker on the trait.
- Formally:

$$\begin{cases} H_0 : P(Y = 1|AA) = P(Y = 1|Aa) = P(Y = 1|aa) \\ H_1 : \text{At least one pair different} \end{cases}$$

- Test statistic

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_2$
- The test makes no assumptions about the relationship between genotype and trait.
- Under H_1, each genotype can have a different disease rate.

Example codominant test

TNF genotype (G/A polymorphism) in a study on acne patients and controls

	GG	AG	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

R code codominant test

```
> X <- matrix(c(66,43,4,99,15,0),byrow=TRUE,ncol=3)
> colnames(X) <- c("GG","GA","AA")
> rownames(X) <- c("Acne","Contro")
> X
       GG GA AA
Acne   66 43  4
Contro 99 15  0

> results <- chisq.test(X)
Warning message:
In chisq.test(X) : Chi-squared approximation may be incorrect
> print(results)

Pearson's Chi-squared test

data:  X
X-squared = 24.113, df = 2, p-value = 5.806e-06

> results$expected
             GG       GA       AA
Acne   82.13656 28.87225 1.991189
Contro 82.86344 29.12775 2.008811
> fisher.test(X)

Fisher's Exact Test for Count Data

data:  X
p-value = 1.97e-06
alternative hypothesis: two.sided
```

Dominant test

- Columns in the original table are combined to produce 2 x 2 tables.
- Dominant model:

	aa	aA or AA	Total
Cases	r_0	$r_1 + r_2$	r
Controls	s_0	$s_1 + s_2$	s
Total	n_0	$n_1 + n_2$	n

- Test

$$\begin{cases} H_0 : \text{Disease does not depend on the presence of A} \\ H_1 : \text{Disease does depend on the presence of A} \end{cases}$$

- Statistic

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_1$

R code dominant test

```
> Y <- cbind(X[,1],X[,2]+X[,3])
> colnames(Y) <- c("GG","GA or AA")
> rownames(Y) <- c("Acne","Control")
> Y
        GG GA or AA
Acne    66       47
Control 99       15

> results <- chisq.test(Y)
> print(results)

Pearson's Chi-squared test with Yates' continuity
correction

data:  Y
X-squared = 21.702, df = 1, p-value = 3.184e-06

> results <- chisq.test(Y,correct=FALSE)
> print(results)

Pearson's Chi-squared test

data:  Y
X-squared = 23.112, df = 1, p-value = 1.528e-06
```

Recessive test

- Recessive model:

	aa or aA	AA	Total
Cases	$r_0 + r_1$	r_2	r
Controls	$s_0 + s_1$	s_2	s
Total	$n_0 + n_1$	n_2	n

- Test

$$\begin{cases} H_0 : \text{Disease does not depend on being homozygote AA} \\ H_1 : \text{Disease does depend on being homozygote AA} \end{cases}$$

- Statistic

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_1$

The additive genetic model

- Basic idea: disease risk increases as a function of the number of alleles (0, 1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - Cochran-Armitage trend test

The alleles test

- Let p be the allele frequency of the A allele.

$$\begin{cases} H_0 : \mathrm{p}_{cases} = p_{controls} \\ H_1 : \mathrm{p}_{cases} \neq p_{controls} \end{cases}$$

- The test assumes Hardy-Weinberg equilibrium
- The test is a χ^2 test for independence in a 2 x 2 table of alleles.

	a	A	Total	$\hat{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/(2r)$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/(2s)$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/(2n)$

Example alleles test

A polymorphism in the Dopamine receptor is supposed to be involved in Schizophrenia. In a case-control study, the following data were obtained:

	11	12	22	Total
Cases	7	69	57	113
Controls	20	56	33	109
Total	27	125	90	242

R code alleles test

```
> X <- matrix(c(7,69,57,20,56,33),byrow=TRUE,ncol=3)
> colnames(X) <- c("11","12","22")
> rownames(X) <- c("Cases","Controls")
> X
         11 12 22
Cases     7 69 57
Controls 20 56 33

> Y <- cbind(2*X[,1]+X[,2],2*X[,3]+X[,2])
> colnames(Y) <- c("1","2")
> Y
          1   2
Cases    83 183
Controls 96 122

> chisq.test(Y,correct=FALSE)

Pearson's Chi-squared test

data:  Y
X-squared = 8.4671, df = 1, p-value = 0.003616

>
```

Cochran-Armitage trend test

- The trend test is based on the linear regression model

$$P(Y = 1|X) = \beta_0 + \beta_1 X + \epsilon$$

- X the number of A alleles (0, 1 or 2)
- Alternatively, we may test equality of means of X ($\bar{X}$) for cases and controls

$$\bar{X}_{cases} = \frac{2r_2 + r_1}{r} = 2\hat{p}_{cases} \quad \bar{X}_{controls} = \frac{2s_2 + s_1}{s} = 2\hat{p}_{controls} \quad \bar{X} = 2\hat{p}$$

- Test:

$$\begin{cases} H_0 : E(X|case) = E(X|controls) \\ H_1 : E(X|case) \neq E(X|controls) \end{cases}$$

- It can be shown that:

$$V(\bar{X}_{cases} - \bar{X}_{controls}) = \frac{4n_2 + n_1 - n\bar{X}^2}{rs}$$

- Then the test statistic for the trend test is

$$Z = \frac{\bar{X}_{cases} - \bar{X}_{controls}}{\sqrt{\frac{4n_2 + n_1 - n\bar{X}^2}{rs}}}$$

- Under H_0; $Z \sim N(0, 1)$. Alternatively, $Z^2 \sim \chi^2_1$

References

Laird, N.M. Lange, C. (2011) *The fundamentals of modern statistical genetics*. Springer.

Computer exercises

A particular SNP is supposed to be involved in Alzheimer's disease. A case control study has been performed, obtaining the following results:

	MM	Mn	nn
Cases	112	278	150
Controls	206	348	150

1. Perform the alleles test for this data set.
2. Perform Cochran-Armitage trend test for this data set.
3. Plot the risk of disease as a function of the number of m alleles. Fit a linear model and add the regression line to the plot. Test the null hypothesis $\beta_1 = 0$.
4. Is there evidence for association of this marker with the disease?
5. Also test for association using a codominant, a dominant and a recessive model.