

Hardy Weinberg and Linkage Equilibrium in Option 11 of Mendel

Chiara Sabatti, UCLA

`csabatti@mednet.ucla.edu`

A Short Course on Statistical Genetics with Mendel
UCLA, September 17-20, 2001

Summary

- **Two notions of equilibrium**
 - Hardy Weinberg
 - Linkage Equilibrium
- **Conditional Exact Tests**
 - Conditionally on the observed alleles frequencies
 - based on exact distributions
- **P-value estimated with random permutations.**

Context

- These notions of equilibrium are often assumed in genetic analysis, one wants to validate them.
- Need a random independent sample of genotypes/haplotypes from the study population (no family type data)

Test of hypothesis

- Genetic equilibrium is considered 'Null Hypothesis'—that is what we assume unless we can prove that it is not reasonable.
- To see how reasonable equilibrium is, we evaluate the probability of recording data as far from equilibrium as the one we observed, if the data was generated under the null hypothesis
- This is formalized in the P-value calculations.

1. Hardy Weinberg Equilibrium

- The hypothesis

$$\Pr(A) = p \quad \Pr(a) = 1 - p$$

AA	aA	aa
p^2	$2p(1 - p)$	$(1 - p)^2$

- The data that we use for the test

AA	aA	aa
n_{AA}	n_{aA}	n_{aa}

The data

Genotype 1	<i>A</i>	<i>A</i>
Genotype 2	<i>A</i>	<i>A</i>
Genotype 3	<i>A</i>	<i>A</i>
Genotype 4	<i>A</i>	<i>a</i>
Genotype 5	<i>A</i>	<i>a</i>
Genotype 6	<i>a</i>	<i>a</i>

⇒

<i>AA</i>	<i>aA</i>	<i>aa</i>
3	2	1

The probability of the data

- Condition on allele counts equal to the observed ones:

$$\begin{array}{c|c} \text{Fr}(A) & 2/3 \\ \hline \text{Fr}(a) & 1/3 \end{array}$$

- Assume there is Hardy-Weinberg equilibrium

$$\begin{array}{c|c} \text{Fr}(AA) & 4/9 \\ \hline \text{Fr}(Aa) & 4/9 \\ \hline \text{Fr}(aa) & 1/9 \end{array}$$

$$\text{Pr}\left(\begin{array}{|c|c|c|} \hline AA & aA & aa \\ \hline 3 & 2 & 1 \\ \hline \end{array} \right) = \frac{6!}{3!2!} \left(\frac{4}{9}\right)^3 \left(\frac{4}{9}\right)^2 \frac{1}{9} = 0.11$$

Fisher exact tests

P-value: probability of recording a result as extreme as the observed one if HW is true (sum of the probabilities of all the possible configurations that have a probability smaller than the one we observed)

⇒ it is not based on asymptotic approximations, but **exact**.

⇒ Calculate the p-value with **permutation**.

Permutations

Genotype 1	<i>A</i>	<i>A</i>
Genotype 2	<i>A</i>	<i>A</i>
Genotype 3	<i>A</i>	<i>A</i>
Genotype 4	<i>A</i>	<i>a</i>
Genotype 5	<i>A</i>	<i>a</i>
Genotype 6	<i>a</i>	<i>a</i>



<i>A</i>	<i>A</i>	<i>A</i>
<i>A</i>	<i>A</i>	<i>A</i>
<i>A</i>	<i>A</i>	<i>a</i>
<i>a</i>	<i>a</i>	<i>a</i>



Genotype 1	<i>a</i>	<i>A</i>
Genotype 2	<i>A</i>	<i>a</i>
Genotype 3	<i>a</i>	<i>A</i>
Genotype 4	<i>A</i>	<i>A</i>
Genotype 5	<i>A</i>	<i>A</i>
Genotype 6	<i>A</i>	<i>a</i>

<i>AA</i>	<i>aA</i>	<i>aa</i>
3	2	1



<i>AA</i>	<i>aA</i>	<i>aa</i>
2	4	0

$$\text{Pr}(\text{obser}) = 0.11$$

$$\text{Pr}(\text{permut}) = 0.46$$

Conclusion

$$\text{P-value} = \frac{\#\text{Permutations : Pr(permut)} \leq \text{Pr(obser)}}{\#\text{Permutations}}$$

⇒ Different from Option 6:

- non parametric;
- not based on asymptotic approximations;
- applicable to context with few data sets.

2. Gametic phase equilibrium

			Alleles' distribution
A1	Marker 1		$A_1 \quad \cdots \quad A_r$ $p_1 \quad \cdots \quad p_r$ <hr/>
B3	Marker 2	(A_1, B_3, C_2)	$B_1 \quad \cdots \quad B_c$ $q_1 \quad \cdots \quad q_c$ <hr/>
		Haplotype	
C2	Marker 3		$C_1 \quad \cdots \quad C_l$ $w_1 \quad \cdots \quad w_l$

Markers 1, 2, and 3 are in equilibrium

$$\Leftrightarrow Pr(A_i B_j C_k) = p_i q_j w_k$$

Linkage equilibrium–Table view

LE = the distribution of the alleles at two (or more) markers is independent. For two markers easy to visualize:

	B_1	B_2	\dots	B_c	
A_1	π_{11}	π_{12}	\dots	π_{1c}	p_1
A_2	π_{21}	π_{22}	\dots	π_{2c}	p_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	π_{r1}	π_{r2}	\dots	π_{rc}	p_r
	q_1	q_2	\dots	q_c	1

=

	B_1	B_2	\dots	B_c	
A_1	$p_1 q_1$	$p_1 q_2$	\dots	$p_1 q_c$	p_1
A_2	$p_2 q_1$	$p_2 q_2$	\dots	$p_2 q_c$	p_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	$p_r q_1$	$p_r q_2$	\dots	$p_r q_c$	p_r
	q_1	q_2	\dots	q_c	1

This may happen in presence of Linkage, Stratification, Epistasis

Why are we interested in it?

- Assumed in many analysis → verify it is appropriate
- Interested in studying the pattern of disequilibrium across the genome and across populations in connection with LD mapping (Option 12)

2.1 The haplotype data

	Marker 1	Marker 2	Marker 3
Haplotype 1	A_1	B_2	C_2
Haplotype 2	A_4	B_4	C_2
Haplotype 3	A_2	B_1	C_4
Haplotype 4	A_3	B_1	C_4
Haplotype 5	A_1	B_3	C_3
⋮	⋮	⋮	⋮
Haplotype n	A_2	B_3	C_1

The data and its probability

(focus on two markers)

table =

	B_1	B_2	\dots	B_c	
A_1	n_{11}	n_{12}	\dots	n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	\dots	n_{2c}	$n_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	n_{r1}	n_{r2}	\dots	n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	\dots	$n_{.c}$	n

- Condition on observed allele frequency
- Assume Linkage Equilibrium

For each table, we can evaluate its probability under the null hypothesis (it's called Fisher-Yates distribution)

$$\Pr(\text{table} | n_{i.}, n_{.j}, LE)$$

Permutation description of null hypothesis

To generate a table from the null distribution I can use permutations:

	Marker 1	Marker 2	Marker 3
Haplotype 1	A_4	B_4	C_4
Haplotype 2	A_1	B_2	C_2
Haplotype 3	A_3	B_1	C_2
Haplotype 4	A_2	B_3	C_4
Haplotype 5	A_1	B_2	C_1
⋮	⋮	⋮	⋮
Haplotype n	A_2	B_1	C_3
	Permut Col 1	Permut Col 2	Permut Col 3

Fisher Exact Test of Independence

P-value: sum of the probabilities of all the tables that have a probability smaller than the one of the observed one.

P-value via permutations

$$\text{P-value} = \frac{\#\text{Permutations : Pr(permut)} \leq \text{Pr(obser)}}{\#\text{Permutations}}$$

⇒ it is not based on asymptotic approximations (as a χ^2 test would be) → good for sparse tables.

⇒ we can estimate the p-value with a random sample of permutations.

2.2 Multilocus Genotype Data

(Phase unknown—for simplicity, biallelic markers)

	Marker 1	Marker 2	Marker 3
Multi-Genotype 1	<i>aa</i>	<i>bB</i>	<i>CC</i>
Multi-Genotype 2	<i>aA</i>	<i>bb</i>	<i>cc</i>
Multi-Genotype 3	<i>aA</i>	<i>BB</i>	<i>cC</i>
Multi-Genotype 4	<i>AA</i>	<i>bB</i>	<i>cC</i>
Multi-Genotype 5	<i>aA</i>	<i>bB</i>	<i>cc</i>
⋮	⋮	⋮	⋮
Multi-Genotype n	<i>aa</i>	<i>BB</i>	<i>CC</i>

Data table and its probability

(focus on two markers)

table =

	<i>bb</i>	<i>bB</i>	<i>BB</i>	
<i>aa</i>	n_{11}	n_{12}	n_{13}	$n_{1.}$
<i>aA</i>	n_{21}	n_{22}	n_{23}	$n_{2.}$
<i>AA</i>	n_{31}	n_{32}	n_{33}	$n_{3.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

- Condition on observed allele frequency
- Assume Linkage Equilibrium
- Assume Hardy Weinberg

For each table, we can evaluate its probability under the null hypothesis

$$\Pr(\text{table} | n_{i.}, n_{.j}, LE)$$

Permutation description of null hypothesis

	Marker 1	Marker 2	Marker 3
Multi-Genotype 1	<i>aA</i>	<i>BB</i>	<i>cC</i>
Multi-Genotype 2	<i>aA</i>	<i>bb</i>	<i>cc</i>
Multi-Genotype 3	<i>aA</i>	<i>bB</i>	<i>CC</i>
Multi-Genotype 4	<i>aA</i>	<i>BB</i>	<i>cC</i>
Multi-Genotype 5	<i>aA</i>	<i>bb</i>	<i>cC</i>
⋮	⋮	⋮	⋮
Multi-Genotype n	<i>aa</i>	<i>BB</i>	<i>cC</i>

↑	↑	↑
<i>AAAAAa</i>	<i>BBBBBB</i>	<i>CCCCCC</i>
<i>aaaaaa...</i>	<i>Bbbbb...</i>	<i>cccccc...</i>

Fisher Exact Test of Independence

NOTE: exactly the same setting as before

P-value: sum of the probabilities of all the tables that have a probability smaller than the one of the observed one.

P-value via permutations

$$\text{P-value} = \frac{\#\text{Permutations} : \Pr(\text{permut}) \leq \Pr(\text{obser})}{\#\text{Permutations}}$$

⇒ it is not based on asymptotic approximations (as a χ^2 test would be) → good for sparse tables.

⇒ we can estimate the p-value with a random sample of permutations.

Outline of input files for Option 11

Locus Standard. Frequency information is not used.

Map Standard. Distance information is not used. Use it to specify on which markers to include in the analysis:

- one marker → Hardy Weinberg
- two or more markers → Linkage Disequilibrium (+ HW)

Pedigree

- They have to be one-person pedigrees.
- You can specify the number of copies of the pedigree.
- One haplotype is entered as a everywhere homozygous multilocus genotype.

Control The required keyword is

OPTION=11

May need to force the program to read the number of copies

READ_PEDIGREE_COPIES=TRUE

To control the number of sampled permutations:

SAMPLE=30000

Option 11 Output files

- The relevant output is the p-value of the conducted test.
- There is also a standard deviation of the p-value.

GENETIC EQUILIBRIUM OPTION

FISHER'S EXACT TEST FOR GENETIC EQUILIBRIUM HAS APPROXIMATE PVALUE 0.4900E-01 PLUS OR MINUS 0.4317E-02 BASED ON 10000 RESAMPLES.

TIME OF OPERATION WAS 7.800000 SECONDS

Other related Mendel options

- A parametric test for HW can be done with Option 6.
- If you are looking for linkage disequilibrium between one known marker and an unknown disease locus, use Option 12.