

Chapter 4. The analysis of Segregation

1 The law of segregation (Mendel's first law)

There are two alleles in the two homologous chromosomes at a certain marker. One of the two alleles received from one of the two alleles of father with equal probability, and likewise one of the two allele from one of the two alleles of mother with equal probability. The segregation analysis is an application of the law of segregation.

- The purpose of segregation analysis is to test segregation ratio (the proportion or probability of the offsprings with certain trait or phenotype).
- If an individual with genotype A_1A_2 has the same phenotype as the individual with genotype A_1A_1 , but has different phenotype from the individual with genotype A_2A_2 , then allele A_1 is said to be dominant to A_2 , or equivalently, allele A_2 is said to be recessive to allele A_1 . In addition, the phenotype associated with A_1 is said to be dominant, and phenotype associated with A_2 is said to be recessive.
- If an individual with genotype A_1A_2 has different phenotype with both A_1A_1 and A_2A_2 , the alleles A_1 and A_2 are said to be codominant.
- For example, there are two alleles D and d at disease locus. Let D be a disease gene (or allele). If the disease is a dominant disease, an individual with genotype DD or Dd will have the disease. An individual with genotype dd will be normal. If the disease is a recessive disease, an individual with genotype DD will have disease, while an individual with genotype Dd or dd will be normal.

2 Segregation analysis for an autosomal dominant disease

Assume that the disease locus has two alleles D and d , and D is the disease allele. If the segregation law is true we can get the following table under the assumption of a dominant disease.

Parents	prob. of genotype			prob. of phenotype	
Mating type	DD	Dd	dd	Affected	Normal
$DD \times DD$	1	0	0	1	0
$DD \times Dd$	0.5	0.5	0	1	0
$DD \times dd$	0	1	0	1	0
$Dd \times Dd$	0.25	0.5	0.25	0.75	0.25
$Dd \times dd$	0	0.5	0.5	0.5	0.5
$dd \times dd$	0	0	1	0	1

Consider a rare disease where the allele D is rare or the allele frequency of allele D is very small. So, majority of the mating type (parents) with possible affected

offspring is $Dd \times dd$. For example, $p_D = 1/1000$. The conditional probabilities of each mating type given that this mating type may have affected offspring are as follow

mating type	$DD \times DD$	$DD \times Dd$	$DD \times dd$	$Dd \times Dd$	$Dd \times dd$
prob.	2.5×10^{-10}	1×10^{-6}	5×10^{-4}	1×10^{-3}	0.9985

So, when we sample parents with one affected and one unaffected, we can suppose that the affected parent has genotype Dd and the unaffected parent has genotype dd . Assume there are total n offspring, r are affected. Now, we will test whether the assumption of the dominant disease is true or not. Let p denote the probability of an offspring being affected. Thus, the test is equivalent to test $H_0 : p = 0.5$.

Let X denote the number of the affected offspring among the n offsprings. Then

$$\chi \sim B(n, p)$$

$$\iff M_2(n, P)$$

where $P = (p_1, p_2)' = (p, 1 - p)$

1. likelihood ratio test. the test statistic ($m = 2$)

$$G^2 = 2 \sum_{i=1}^m X_i \log \frac{X_i}{np_i^0} = 2(X \log \frac{2X}{n} + (n - X) \log \frac{2(n - X)}{n})$$

2. Score test or Pearsons' chi-square test statistic

$$\begin{aligned} S^2 &= \sum_{i=1}^m \frac{[X_i - np_i^0]^2}{np_i^0} = \frac{[X - n/2]^2}{n/2} + \frac{[n - X - n/2]^2}{n/2} \\ &= \frac{4}{n} [X - n/2]^2 \end{aligned}$$

- Example: A study on opalescent dentine examined a random sample of 112 offspring of the matings between affected and unaffected individuals, and found that 52 were affected while the other 60 were normal. Are these data consistent with hypothesis that opalescent dentine is a rare autosomal dominant disease?

Let p be the segregation rate. In fact, we want to test the null hypothesis $H_0 : p = 0.5$. Using the three test statistics given above, we can calculate, using $n = 112, X = 52$,

$$G^2(x) = 0.5719$$

$$S^2(x) = 0.5714$$

The p-value of the three tests,

$$p_{G^2} = P(\chi_1^2 > 0.5719) = 0.4497$$

$$p_{S^2} = P(\chi_1^2 > 0.5714) = 0.4497$$

So, none of the two tests can reject the null hypothesis i.e. we can not reject the statement "opalescent dentine is a rare autosomal dominant disease".

For the locus with codominant alleles, we can use the similar methods to construct the three test statistics. For example, MN blood type, every individual can be classified into three groups: MM, MN and NN (observable) of phenotypes also genotypes. The data of n offspring of mating type $MN \times MN$ can be summarized as

Phenotype	MM	MN	NN
number of individuals	n_1	n_2	n_3

So, we can construct the three tests with multinomial distribution $M_3(n, p)$. We want to test $H_0 : p_1 = 1/4; p_2 = 1/2$ and $p_3 = 1/4$. For recessive disease, test of segregation will be more complicated. We will not give the details here.

3 Interpreting the deviation from Mendelian segregation ratio

The Mendelian segregation is true for the disease that is determined by the alleles at a single locus. The deviation from the Mendelian segregation rate may be because

1. The trait (phenotype) is not determined by a single locus.
2. The trait or disease due to the mixture of the genetic and environmental factors.
3. Incomplete penetrance

More general disease model: Let D and d denote the two alleles of the disease model. The probabilities,

$$f_{DD} = P(Affected|DD)$$

$$f_{Dd} = P(Affected|Dd)$$

$$f_{dd} = P(Affected|dd)$$

called penetrances, may not be 0 or 1. For example, $f_{DD} = 0.5, f_{Dd} = 0.2, f_{dd} = 0.1$. If we assume D is high risk allele, then

$$f_{DD} \geq f_{Dd} \geq f_{dd}$$

- If $f_{DD} = f_{Dd} \neq f_{dd}$, the disease is called dominant disease and the model is called dominant disease model.
- If $f_{DD} \neq f_{Dd} = f_{dd}$, the disease is called recessive disease and the model is called recessive disease model.
- If $f_{DD} = \frac{1}{2}(f_{Dd} + f_{dd})$, the model is called additive disease model.
- If $f_{DD} = \sqrt{f_{Dd}f_{dd}}$, the model is called multiplicative disease model.

4 Hardy-weinberg equilibrium

For a biallelic marker with allele A and a , let P_{AA} , P_{Aa} and P_{aa} denote the genotype frequencies of the three genotype AA , Aa , and aa , respectively. let p_A denote the allele frequency of allele A ($1 - p_A$ will be the allele frequency of allele a).

- Hardy-Weinberg Equilibrium: Under random mating

$$\begin{aligned} P_{AA} &= p_A^2. \\ P_{Aa} &= 2p_A(1 - p_A). \\ P_{aa} &= (1 - p_A)^2. \end{aligned}$$

Why this is true and why called equilibrium? Assume that the individuals of the i th generation are the offspring through random mating of the $(i - 1)$ th generation ($i = 1, 2, \dots$). Denote the genotype frequencies at the i th generation by $P_{AA}^{(i)}$, $P_{Aa}^{(i)}$, and $P_{aa}^{(i)}$. ($i = 0, 1, 2, \dots$). and the allele frequency of allele A by $P^{(i)}$. Then, using the table,

Parents Mating type	prob. of genotype of offspring		
	AA	Aa	aa
$AA \times AA$ (M1)	1	0	0
$AA \times Aa$ (M2)	0.5	0.5	0
$AA \times aa$ (M3)	0	1	0
$Aa \times Aa$ (M4)	0.25	0.5	0.25
$Aa \times aa$ (M5)	0	0.5	0.5
$aa \times aa$ (M6)	0	0	1

we have, the genotype frequencies in the 1st generation,

$$\begin{aligned} P_{AA}^{(1)} &= P(AA) = \sum_{i=1}^6 P(AA|M_i)P(M_i) \\ &= 1 \cdot P(AA \times AA) + \frac{1}{2} \cdot P(AA \times Aa) + \frac{1}{4} \cdot P(Aa \times Aa) \\ &= (P_{AA}^{(0)})^2 + \frac{1}{2}P_{AA}^{(0)}P_{Aa}^{(0)} + \frac{1}{4}(P_{Aa}^{(0)})^2 \\ &= (P_{AA}^{(0)} + \frac{1}{2}P_{Aa}^{(0)})^2 \\ &= (P(AA) + P(Aa \text{ with order}))^2 = (p^{(0)})^2 \end{aligned} \tag{1}$$

Similarly,

$$P_{Aa}^{(2)} = 2(P_{AA}^{(0)} + \frac{1}{2}P_{Aa}^{(0)})(P_{aa}^{(0)} + \frac{1}{2}P_{Aa}^{(0)}) = 2p^{(0)}(1 - p^{(0)})$$

$$P_{aa}^{(1)} = (P(aa) + P(Aa \text{ with order}))^2 = (1 - p^{(0)})^2.$$

Similarly, we can get the frequencies of the second generation

$$P_{AA}^{(1)} = (P_{AA}^{(1)} + \frac{1}{2}P_{Aa}^{(1)})^2$$

$$= \begin{cases} (p^{(1)})^2 & \text{and} \\ [(p^{(0)})^2 + p^{(0)}(1 - p^{(0)})]^2 = [p^{(0)}]^2 \end{cases}$$

Therefore $p^{(0)} = p^{(1)}$ and $P_{AA}^{(1)} = P_{AA}^{(2)}$. In the same way, we can get $P_{Aa}^{(1)} = P_{Aa}^{(2)}$, and $P_{aa}^{(1)} = P_{aa}^{(2)}$. That means that the frequencies of the genotypes and the allele frequency remain the same after first generation. So, it is called equilibrium.

Note that the genotype frequencies of generation 0 are not necessarily the same as that of the first generation. For example, 0 generation is a mixture of two sub-populations. The individuals of the first generation consist 30% of individuals from ancestral population 1 with genotype AA and 70% of the individuals from ancestral population 2 with genotype aa . So, $P_{AA}^{(0)} = 0.3, P_{aa}^{(0)} = 0.7, P_{Aa}^{(0)} = 0$ and $p^{(0)} = 0.3$ (at generation 0, $P_{AA} = (p_A)^2, P_{Aa} = 2p_A(1 - p_A)$ and $P_{aa} = (1 - p_A)^2$ are not true)

In fact, when we deduce the formula, we need, except the assumption of random mating, the following conditions: (1) infinite population size, (2) discrete generations, (3) no selection, (4) no migration, (5) no mutation and (6) equal initial genotype frequencies in the two sexes.

For the multi-allelic marker with alleles A_1, A_2, \dots, A_L for genotype $G_{ij} = A_i A_j$, the Hardy-Weinberg equilibrium means that

$$P(G_{ij}) = \begin{cases} P(A_i)P(A_j) & i \neq j \\ P^2(A_i) & i = j. \end{cases}$$

for any $1 \leq i \leq j \leq L$.

- Test Hardy-Weinburg equilibrium

Consider a marker with L codominant alleles A_1, A_2, \dots, A_L . Use G_{ij} to denote the genotype $A_i A_j$; $1 \leq i \leq j \leq L$. We sampled n individuals. Let Y_{ij} denote the number of individuals with genotype G_{ij} ($\sum_{1 \leq i \leq j \leq L} Y_{ij} = n$) and

$$Y = (Y_{11}, Y_{12}, \dots, Y_{1L}, Y_{22}, \dots, Y_{2L}, \dots, Y_{LL})'$$

has multinomial distribution $M_m(n, P)$ where $m = L(L+1)/2$. $P = (P_{11}, P_{12}, \dots, P_{1L}, P_{22}, \dots, P_{2L}, \dots, P_{LL})'$. Furthermore, let p_i denote the allele frequency of allele A_i . Testing Hardy-Weinberg equilibrium is equivalent to test

$$H_0 : P_{ij} = \delta_{ij} p_i p_j,$$

where $\delta_{ij} = 2$ if $i \neq j$, and $\delta_{ij} = 1$ if $i = j$.

- Hardy-Weinberg equilibrium for multi-marker haplotype

Let H_1 and H_2 be two multi-marker haplotypes. H_1H_2 is a multimarker genotype. In that case, Hardy-Weinberg equilibrium means, for any two haplotypes H_1 and H_2

$$P(H_1H_2) = \begin{cases} 2P(H_1)P(H_2) & H_1 \neq H_2 \\ P^2(H_1) & H_1 = H_2 \end{cases}$$