The Study of Conditional Probability Matrix for Realive Pair Genotypes

Chaobing He

School of Mathematics and Statistics, Anyang Normal University, Anyang 455000, China

Email: chaobing5@163.com

Abstract: This paper considers the conditional probability matrix for realive pair genotypes. First this paper systematically introduces Hardy-Weinberg equilibrium law, the joint probability distributions for relative pair genotypes and the ITO method. Then this paper discusses the calculation of probability distribution of the number of identical-by-descent allele of relative pairs, and lists the probabilities in tables. The paper gives the definition of quasi conditional probability matrix for the realive pair genotypes, and obtains a conclusion about quasi conditional probability matrix.

Keywords: quasi conditional probability matrix; ITO method; relative pair; genotype; Hardy-Weinberg equilibrium law

Mathematics Subject Classification: 62P10

1. Introduction

A kind of calculation method named ITO method came on (see [1]) to calculate the joint probability distribution for relative pair genotypes. With the ITO method, given the genotype of an individual, it is possible to derive the conditional probability of the genotypes of any non-inbred relative of that individual. The ITO method was extended to handle multiple alleles and was generalized for inbred populations (see [2]). The ITO method was generalized for multiple loci and was also extended to handle consanguinity (see [3]). [4] extended the ITO method to handle ordered genotypes. [5] gived an exact calculation of the probability of identity-by-descent in two-locus models using an extension of the Li-Sacks' method. Studies on the distribution of relative pair genotypes are available in the literature [6–13]. It is very important to study the conditional probability matrix for realive pair genotypes in the ITO method.

The rest of this paper is organized as follows. In Section 2, we describe Hardy-Weinberg equilibrium law. In Section 3, ITO method is introduced in detail, and we consider the joint probability distribution matrix for relative pair genotypes. In Section 4, we discuss the calculation of probability distribution of the number of identical-by-descent allele of relative pairs. In Section 5, we give the definition of quasi conditional probability matrix for the realive pair genotypes, and obtain a conclusion about quasi conditional probability matrix. Finally, we summarize and conclude the paper in Section 6.

2. Hardy-Weinberg equilibrium law

British mathematician Hardy [14] and German physiologist Weinberg [15] published the equilibrium law in the genetics at the same time in 1908. Hardy-Weinberg equilibrium law is derived under the assumption of random mating and the principle of independent segregation. Random mating means that any woman is equally likely to marry any man. The principle of independent segregation is that a mother (or father) is equally likely to pass on either of the two alleles to her offspring (both are 1/2), and that maternal and paternal alleles are inherited independently. Considering that there are two alleles A and a in a locus, it is assumed that the probabilities of the two alleles in the population of parental generation are equal to

$$P(A) = p, \quad P(a) = 1 - p = q. \tag{1}$$

If the relationship between the probabilities of three genotypes and the probabilities of alleles in a population is as follows:

$$P(AA) = p^2, \quad P(Aa) = 2pq, \quad P(aa) = q^2, \tag{2}$$

the genotype probabilities of this population at this locus are said to have the Hardy-Weinberg proportion.

3. Joint probability distribution for relative pair genotypes

In order to calculate the joint probability distribution for general relative pair genotypes, a kind of mechanized calculation method named ITO method was proposed in literature [1]. ITO method is as follows. R_1 and R_2 are used to represent genotypes of the relative pairs at the given locus respectively. Assuming that there are two alleles A and a at this locus, their probabilities are p and q, respectively. If the numbers 0, 1 and 2 represent three genotypes a, Aa and AA, then joint probability distribution with respect to the genotypes of relative pairs is

$$P(R_1 = i, R_2 = j) = P(R_1 = i|R_2 = j)P(R_2 = j), i, j = 0, 1, 2,$$

where the marginal probability $P(R_2 = j) = C_2^j p^j (1-p)^{2-j}$ is easily calculated by the Hardy-Weinberg equilibrium law.

The conditional probability $P(R_1 = i|R_2 = j)$ is calculated as follows. Let IBD denote the number of identical-by-descent allele of relative pairs, which is a random variable with the values 0,1 and 2. By the total probability formula,

$$\begin{aligned} P(R_1 = i|R_2 = j) &= \sum_{t=0}^{2} P(R_1 = i|IBD = t, R_2 = j)P(IBD = t|R_2 = j) \\ &= \sum_{t=0}^{2} P(R_1 = i|IBD = t, R_2 = j)P(IBD = t). \end{aligned}$$

For $IBD = 0, 1, 2$, a matrix can be used to represent the value of genotype conditional probability $p_{ij} = P(R_1 = i|IBD = t, R_2 = j)$.

When $BID = 2$, the conditional probability p_{ij} is given by following matrix:

$$\begin{array}{cc} & \begin{array}{ccc} R_1{=}AA & Aa & aa \end{array} \quad given\ R_2{=} \\ I = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \tag{3}$$

When $BID = 1$, the conditional probability p_{ij} is given by following matrix:

$$\begin{array}{cc} & \begin{array}{ccc} R_1{=}AA & Aa & aa \end{array} \quad given\ R_2{=} \\ T = & \begin{pmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{pmatrix} \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \tag{4}$$

When $BID = 0$, the conditional probability p_{ij} is given by following matrix:

$$\begin{array}{c c} \begin{array}{cccc} R_1{=}AA & Aa & aa & given\ R_2{=} \end{array} & \\ O = \begin{pmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{pmatrix} & \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \tag{5}$$

Set $\Delta_i = P(IBD = i), i = 0, 1, 2$. The conditional probability matrix (CPM) for relative pair genotypes is given by

$$W = \Delta_2 I + \Delta_1 T + \Delta_0 O. \tag{6}$$

The joint probability distribution matrix for relative pair genotypes is

$$C = \begin{pmatrix} p^2 & 0 & 0 \\ 0 & 2pq & 0 \\ 0 & 0 & q^2 \end{pmatrix} W \tag{7}$$

So it's important to study Δ_2, Δ_1 and Δ_0.

4. Calculation of Δ_2, Δ_1 and Δ_0

Suppose the conditional probability matrix for the realive pair genotypes is

$$W = \Delta_2 I + \Delta_1 T + \Delta_0 O, \Delta_2 \neq 0. \tag{8}$$

So let's discuss the calculation of Δ_2, Δ_1 and Δ_0. For convenience, assume that the fathers of the realive pair have the same ancestors, and the mothers for the realive pair have the same ancestors.

Suppose the conditional probability matrix for the fathers is

$$W_1 = \Delta_{21} I + \Delta_{11} T + \Delta_{01} O. \tag{9}$$

Suppose the conditional probability matrix for the mothers is

$$W_2 = \Delta_{22} I + \Delta_{12} T + \Delta_{02} O. \tag{10}$$

suppose $A_1 =$"the realive pair share a gene identical by descent with each other, respectively from their own father".

suppose $A_2 =$"the realive pair share a gene identical by descent with each other, respectively from their own mother".

One has

$$P(A_1) = \frac{1}{4}\Delta_{11} + \frac{1}{2}\Delta_{21} \triangleq \Phi_1, \tag{11}$$

$$P(A_2) = \frac{1}{4}\Delta_{12} + \frac{1}{2}\Delta_{22} \triangleq \Phi_2. \tag{12}$$

In fact, Φ_1 and Φ_2 are both the coefficients of relationship.

Hence,

$$\Delta_2 = \Phi_1\Phi_2,$$

$$\Delta_1 = \Phi_1(1-\Phi_2) + \Phi_2(1-\Phi_1),$$

$$\Delta_0 = (1-\Phi_1)(1-\Phi_2).$$

Set vector $\Delta = (\Delta_2, \Delta_1, \Delta_0)$.

The first $\Delta = (1, 0, 0)$.

The second Δ is obtained as follows:

According to the first $\Delta = (1, 0, 0)$, we obtain the coefficients of relationship

$$\Phi = \frac{1}{4}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2}.$$

Hence,

$$\Delta_2 = \Phi\Phi = \frac{1}{4},$$

$$\Delta_1 = \Phi(1-\Phi) + \Phi(1-\Phi) = \frac{1}{2},$$

$$\Delta_0 = (1-\Phi)(1-\Phi) = \frac{1}{4}.$$

Hence, The second $\Delta = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

The third Δ is obtained as follows:

According to the second $\Delta = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, we obtain the coefficients of relationship

$$\Phi = \frac{1}{4}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{4}.$$

Hence,

$$\Delta_2 = \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8},$$

$$\Delta_1 = \frac{1}{2}\cdot(1-\frac{1}{4}) + (1-\frac{1}{2})\cdot\frac{1}{4} = \frac{1}{8} = \frac{4}{8},$$

$$\Delta_0 = (1-\frac{1}{2})(1-\frac{1}{4}) = \frac{3}{8}.$$

Hence, The third $\Delta = (\frac{1}{8}, \frac{4}{8}, \frac{3}{8})$.

The rest of Δs can be done in the same manner as before.

Table 1 and Table 2 show the Δs according to the Φs

$$1/2, 1/4, 3/16, 11/64, 5/32, 1/8, 7/64, 3/32, 5/64, 1/16.$$

If the conditional probability matrix for the realive pair genotypes is

$$W = \Delta_1 T + \Delta_0 O, \Delta_1 \neq 0, \tag{13}$$

then $\Delta_1 = \Phi$ in Table 1 and Table 2.

5. Quasi conditional probability matrix for realive pair genotypes

Definition 1 *$W = aI + bT + cO$ is called a quasi conditional probability matrix (QCPM) for realive pair genotypes if*

(1) (non-negativity) a, b, c are all nonnegative rational number;

(2) (normativity) $a + b + c = 1$;

(3) (decomposability) there exist real numbers x, y such that $a = xy$ and $c = (1 - x)(1 - y)$.

In fact, (3) in Definition 1 can be replaced by the following

(3') $(b + 2a)^2 \geq 4a$.

The equivalence proof of (3) and (3') is as follows:

Proof: (3)$\Longrightarrow$(3')

From (3), we get $a = xy, c = 1 - (x + y) + xy$.

Hence, $xy = a, x + y = 1 + a - c = b + 2a$.

The real numbers x, y are roots of the fillowing equation with respect to z,

$$z^2 - (b + 2a)z + a = 0. \tag{14}$$

Obviously, $(b + 2a)^2 - 4a \geq 0$.

(3')$\Longrightarrow$(3)

Since $(b+2a)^2 \geq 4a$, let x, y be roots of the fillowing equation with respect to z,

$$z^2 - (b+2a)z + a = 0. \tag{15}$$

Obviously, $x + y = b + 2a = 1 + a - c, xy = a$.

Hence, $a = xy, c = 1 - (x+y) + xy = (1-x)(1-y)$.

Obviously, CPM is a QCPM.

Lemma 1 *Suppose* $a + b + c = 1, a \geq 0, b \geq 0, c \geq 0$. *Then the following three inequalities are equivalent*

(1) $4a \leq (b+2a)^2$*;*

(2) $(b+2a)^2 \leq \frac{b^2}{c}$*;*

(3) $4a \leq \frac{b^2}{c}$.

Proof: (1)$\Longrightarrow$(2)

Since $(b+2a)^2 \geq 4a$, let x, y be roots of the fillowing equation with respect to z,

$$z^2 - (b+2a)z + a = 0, \tag{16}$$

so, $x + y = b + 2a, xy = a$.

Hence, $a = xy, b = x + y - 2xy, c = 1 - x - y + xy$.

Then

$$\begin{aligned}
& b^2 - c(b+2a)^2 \\
&= (x+y-2xy)^2 - (1-x-y+xy)(x+y)^2 \\
&= [(x+y)^2 - 4xy(x+y) + 4x^2y^2] \\
&\quad - [(x+y)^2 - (x+y)^3 + xy(x+y)^2] \\
&= [(x+y)^3 - 4xy(x+y)] \\
&\quad - [xy(x+y)^2 - (x+y)^3 - 4x^2y^2] \\
&= (x+y)(x-y)^2 - xy(x-y)^2 \\
&= (x-y)^2(x+y-xy) \\
&= (x-y)^2(b+a) \geq 0,
\end{aligned}$$

i.e.,

$$(b+2a)^2 \leq \frac{b^2}{c}. \tag{17}$$

(2)$\Longrightarrow$(1)

Since

$$\begin{aligned} & b^2 - c(b+2a)^2 \\ =& b^2 - (1-a-b)(b+2a)^2 \\ =& (a+b)[4a(a+b) - 4a + b^2] \\ =& (a+b)[b^2 + 4ab + 4(a^2 - a)] \geq 0, \end{aligned}$$

hence $b \geq -2a + 2\sqrt{a}$, i.e.,

$$(b+2a)^2 \geq 4a.$$

(1)$\Longrightarrow$(3)

Since (1)$\Longrightarrow$(2), so from (1) and (2), (3) holds clearly.

(3)$\Longrightarrow$(1)

Since

$$\begin{aligned} & b^2 - 4ac \\ =& b^2 - 4a(1-a-b) \\ =& b^2 + 4ab + 4(a^2 - a) \geq 0, \end{aligned}$$

hence $b \geq -2a + 2\sqrt{a}$, i.e., $(b+2a)^2 \geq 4a$.

Conclusion 1 *If W is a quasi conditional probability matrix for the realive pair genotypes, then*

$$4a \leq (b+2a)^2 \leq \frac{b^2}{c}. \tag{18}$$

According to Definition 1 and Lemma 1, Conclusion 1 is easy to be proved.

6. Conclusions

In this paper, we consider the conditional probability matrix for realive pair genotypes. We discuss the calculation of probability distribution of the number

Table 1 The first 30 sets of values of $\triangle_2, \triangle_1, \triangle_0, \Phi_1, \Phi_2, \Phi$

No	$\triangle_2$	$\triangle_1$	$\triangle_0$	Φ_1	Φ_2	Φ
1	1	0	0	1	1	1/2
2	1/4	1/2	1/4	1/2	1/2	1/4
3	1/8	1/2	3/8	1/2	1/4	3/16
4	3/32	1/2	13/32	1/2	3/16	11/64
5	11/128	1/2	53/128	1/2	11/64	43/256
6	5/64	1/2	27/64	1/2	5/32	21/128
7	1/16	1/2	7/16	1/2	1/8	5/32
8	1/16	3/8	9/16	1/4	1/4	1/8
9	7/128	1/2	57/128	1/2	7/64	39/256
10	3/64	1/2	29/64	1/2	3/32	19/128
11	3/64	11/32	39/64	1/4	3/16	7/64
12	11/256	43/128	159/256	1/4	11/64	27/256
13	5/128	1/2	59/128	1/2	5/64	37/256
14	5/128	21/64	81/128	1/4	5/32	13/128
15	9/256	39/128	169/256	3/16	3/16	3/32
16	33/1024	151/512	689/1024	3/16	11/64	23/256
17	1/32	1/2	15/32	1/2	1/16	9/64
18	1/32	5/16	21/32	1/4	1/8	3/32
19	121/4096	583/2048	2809/4096	11/64	11/64	11/128
20	15/512	73/256	351/512	3/16	5/32	11/128
21	7/256	39/128	171/256	1/4	7/64	23/256
22	55/2048	281/1024	1431/2048	11/64	5/32	21/256
23	25/1024	135/512	729/1024	5/32	5/32	5/64
24	3/128	19/64	87/128	1/4	3/32	11/128
25	3/128	17/64	91/128	3/16	1/8	5/64
26	11/512	65/256	371/512	1/8	11/64	19/256
27	21/1024	131/512	144/199	3/16	7/64	19/256
28	5/256	37/128	177/256	1/4	5/64	21/256
29	5/256	31/128	189/256	1/8	5/32	9/128
30	77/4096	499/2048	1703/2309	11/64	7/64	9/128

Table 2 Other sets of values of $\triangle_2, \triangle_1, \triangle_0, \Phi_1, \Phi_2, \Phi$

No	$\triangle_2$	$\triangle_1$	$\triangle_0$	Φ_1	Φ_2	Φ
31	9/512	63/256	377/512	3/16	3/32	9/128
32	35/2048	237/1024	1539/2048	5/32	7/64	17/256
33	33/2048	239/1024	1537/2048	11/64	3/32	17/256
34	1/64	9/32	45/64	1/4	1/16	5/64
35	1/64	7/32	49/64	1/8	1/8	1/16
36	15/1024	121/512	767/1024	3/16	5/64	17/256
37	15/1024	113/512	783/1024	5/32	3/32	1/16
38	7/512	53/256	399/512	1/8	7/64	15/256
39	55/4096	457/2048	3127/4096	11/64	5/64	1/16
40	25/2048	215/1024	1593/2048	5/32	5/64	15/256
41	49/4096	399/2048	3249/4096	7/64	7/64	7/128
42	3/256	29/128	195/256	3/16	1/16	1/16
43	3/256	25/128	203/256	1/8	3/32	7/128
44	11/1024	109/512	795/1024	11/64	1/16	15/256
45	21/2048	187/1024	1653/2048	7/64	3/32	13/256
46	5/512	47/256	413/512	1/8	5/64	13/256
47	5/512	51/256	405/512	5/32	1/16	7/128
48	9/1024	87/512	841/1024	3/32	3/32	3/64
49	35/4096	349/2048	3363/4096	7/64	5/64	3/64
50	1/128	11/64	105/128	1/8	1/16	3/64
51	15/2048	161/1024	1711/2048	3/32	5/64	11/256
52	7/1024	81/512	855/1024	7/64	1/16	11/256
53	25/4096	295/2048	3481/4096	5/64	5/64	5/128
54	3/512	37/256	435/512	3/32	1/16	5/128
55	5/1024	67/512	885/1024	5/64	1/16	9/256
56	1/256	15/128	225/256	1/16	1/16	1/32

of identical-by-descent allele of relative pairs, and list the probabilities in tables. We give the definition of quasi conditional probability matrix for the realive pair genotypes, and obtain a conclusion about quasi conditional probability matrix.

Reference

[1] C. C. Li, L. Sacks, The derivation of joint distribution and correlation between relatives by the use of stochastic matrices, Biometrics 10 (3) (1954) 347–360.

[2] W. H. Richardson, Frequencies of genotypes of relatives, as determined by stochastic matrices, Genetica 35 (1964) 323–354.

[3] E. R. C. Campbell M A, Relatives of probands: models for preliminary genetic analysis, Annals of human genetics 35 (1971) 225–236.

[4] D. Feng, D. E. Weeks, Ordered genotypes: An extended ito method and a general formula for genetic covariance, American Journal of Human Genetics 78 (6) (2006) 1035–1045.

[5] W. Li, An exact calculation of the probability of identity-by-descent in two-locus models using an extension of the li-sacks’ method, American journal of human genetics 63 (1998) A297.

[6] M. P. Epstein, W. L. Duren, M. Boehnke, Improved inference of relationship for pairs of individuals, The American Journal of Human Genetics 67 (5) (2000) 1219–1231.

[7] R. S. Legro, R. Spielman, M. Urbanek, D. Driscoll, J. F. Strauss, A. Dunaif, Phenotype and genotype in polycystic ovary syndrome, Recent progress in hormone research 53 (1998) 217–256.

[8] D. O. Stram, C. L. Pearce, P. Bretsky, M. Freedman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel, B. E. Henderson, D. C. Thomas, Modeling and em estimation of haplotype-specific relative risks from genotype data

for a case-control study of unrelated individuals, Human heredity 55 (4) (2003) 179–190.

[9] M. Boehnke, N. J. Cox, Accurate inference of relationships in sib-pair linkage studies, The American Journal of Human Genetics 61 (2) (1997) 423–429.

[10] B. A. Rybicki, R. C. Elston, The relationship between the sibling recurrence-risk ratio and genotype relative risk, The American Journal of Human Genetics 66 (2) (2000) 593–604.

[11] D. L. Harris, Genotypic covariances between inbred relatives., Genetics 50 (1964) 1319–1348.

[12] M. Crabtree, I. Tomlinson, S. Hodgson, K. Neale, R. Phillips, R. Houlston, Explaining variation in familial adenomatous polyposis: relationship between genotype and phenotype and evidence for modifier genes, Gut 51 (3) (2002) 420–423.

[13] M. R. Robinson, G. English, G. Moser, L. R. Lloyd-Jones, M. A. Triplett, Z. Zhu, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, et al., Genotype–covariate interaction effects and the heritability of adult body mass index, Nature genetics 49 (8) (2017) 1174–1181.

[14] G. H. Hardy, Mendelian proportions in a mixed population, Science 28 (1954) 49–50.

[15] W. Weinberg, Über den nachweis der vererbung beim menschen, Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg 64 (1908) 368–382.