EPH - International Journal of Mathematics and Statistics

ISSN (Online): 2208-2212
Volume 2 Issue 1 June 2016

DOI:https://doi.org/10.53555/eijms.v5i1.39

RANDOM SIMULATION OF GENOTYPES FROM RELATIVE PAIRS USING R

Chaobing He[1*]
*[1]*School of Mathematics and Statistics, Anyang Normal University, Anyang 455000, China*

Corresponding Author:-

Abstract:-

This paper considers the random simulation of genotypes from relative pairs using R. After Hardy-Weinberg equilibrium law is introduced systematically, the joint probability distributions of general relative pairs are studie. Then ITO method was discussed in detail. Based on the theo ry described above, computer program for genotypes from relative pairs is written using R. According to these R codes, genotypes from relative pairs can be easily generated randomly.

Keywords:-*Relative pair; genotype; Hardy-Weinberg equilibrium law; ITO method; conditional probability matrix*

Mathematics Subject Classification: *62P10; 65C05; 65C10*

1. INTRODUCTION

In order to calculate the joint probability distribution of general relative pairs, a kind of mechanized calculation method named ITO method was proposed (see [1]). 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: $P(R_2|R_1)$, where R_i denotes the genotype of the ith person. 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. For research purposes, we often use software to randomly generate genotypes from relative pairs. Therefore, it is very important to write computer programs to simulate genotypes from relative pairs.

This work was supported in part by the Key Scientific Research Project of Colleges and Universities of Henan Province of China under Grant 16A110001 and Grant 18A110009.

The rest of this paper is organized as follows. In Section 2, we present a description of Hardy-Weinberg equilibrium law. In Section 3, the joint probability distributions of general relative pairs are studie, then ITO method is discussed in detail. In Section 4, computer program for genotypes from relative pairs is written using R.

2. Hardy-Weinberg equilibrium law

British mathematician Hardy [6] and German physiologist Weinberg [7] published the equilibrium law in the genetics at the same time in 1908. They proved that genotype probability reach equilibrium after the generation of random mating, and keep the equilibrium later, unless population allele probability is changed by some factors. Hardy-Weinberg equilibrium law plays an important role in the study of genetics.

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

Now let's introduce the Hardy-Weinberg equilibrium law. 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

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

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

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

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

If the genotype probabilities of parental generation have the Hardey-Weinberg proportion (2), the genotype probabilities of offspring generation will have the Hardy-Weinberg proportion under the assumption of random parental mating.

If the genotype probability of parental generation does not have the HardeyWeinberg proportion, the genotype probability of the offspring generation will have the Hardey-Weinberg proportion under the assumption of random parental mating.

3. Joint probability distribution about the genotypes of relative pairs

Relatives refer to people with blood relationship, and blood relationship in genetics means that the genes may be of the same origin, that is, they have the same ancestors. Because the number of individuals in a population is limited, if the ancestors of any two people are traced, their common ancestor will always be found in a certain generation, that is, they will always have blood relations. The blood relationship discussed in genetics is relative and generally refers to relative within three generations. A couple, if they are not close relatives, are said that they are not relatives. Father and son, brother and sister, and grandparent and grandchild are relatives. Since relatives may share alleles, their genotypes are not independent. The ITO method provides an elegant algorithm for deriving joint genotype probabilities between pairs of relatives. Now let's introduce the ITO method in detail. 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

$$P(R_1 = i, R_2 = j) = P(R_1 = i|R_2 = j)P(R_2 = j), i,j = 0,1,2, \tag{3}$$

where the marginal probability$P(R_2 = j) = C_2^j p^j (1-p)^{2-j}$ is easily calculate

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,

$$\begin{aligned} P(R_1 = i|R_2 = j) &= \sum_{t=0}^{2} P(R_1 = i|IBD = t, R_2 = j)P(IBD = t|R_2 = j) \\ &= \sum_{t=0}^{2} P(R_1 = i|IBD = t, R_2 = j)P(IBD = t). \end{aligned} \tag{4}$$

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:

$$\begin{array}{rcl} & \begin{array}{cccc} R_1{=}AA & Aa & aa & \quad given\ R_2{=} \end{array} & \\ I = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \tag{5}$$

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

$$\begin{array}{rcl} & \begin{array}{cccc} R_1{=}AA & Aa & aa & \quad given\ R_2{=} \end{array} & \\ T = & \begin{pmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{pmatrix} & \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \tag{6}$$

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

$$\begin{array}{rcl} & \begin{array}{cccc} R_1{=}AA & Aa & aa & \quad given\ R_2{=} \end{array} & \\ O = & \begin{pmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{pmatrix} & \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \tag{7}$$

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

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

The joint probability distribution matrix of relative pair genotype

$$C = \begin{pmatrix} p^2 & 0 & 0 \\ 0 & 2pq & 0 \\ 0 & 0 & q^2 \end{pmatrix} W \tag{9}$$

The recurrence formula for T^n is given by

$$T^{n+1} = \left(\frac{1}{2}\right)^n T + \left[1 - \left(\frac{1}{2}\right)^n\right] O, \tag{10}$$

where $n + 1$ is the number of generations between the two relatives.

As a special case,

$$T^2 = \frac{1}{2}T + \frac{1}{2}O, \tag{11}$$

and T^2 gives the conditional probabilities for a grandparent-grandchild pair, or half-sibs, who have one parent in common. From the Equation (11), it follows that

$$O = 2T^2 - T. \tag{12}$$

After the derivation of $\Delta_i s$, the conditional probability matrix for full sibs is

$$S = \frac{1}{4}I + \frac{1}{2}T + \frac{1}{4}O. \tag{13}$$

From the Equation (13), it follows that

$$S^2 = \left(\frac{1}{4}I + \frac{1}{2}T + \frac{1}{4}O\right) S \tag{14}$$

$$= \frac{1}{4}S + \frac{1}{2}TS + \frac{1}{4}OS. \tag{15}$$

Hence,

$$OS = 4S^2 - S - 2T^2. \tag{16}$$

In fact, the conditional probability matrix for double first cousins, whose parents are members of two sibships, is

$$S^2 = \frac{1}{16}I + \frac{6}{16}T + \frac{9}{16}O. \tag{17}$$

The matrixes T and S have the following properties:

$$ST = TS = T^2, \qquad TST = STS = T^3. \tag{18}$$

$ST = TS = T^2$ gives the conditional probabilities for a uncle-nephew pair.
Now summarize the vectors $\Delta = (\Delta_2, \Delta_1, \Delta_0)$ for common relative pairs:
(1) Identical twins: $\Delta = (1,0,0)$;
(2) Father and son: $\Delta = (0,1,0)$;
(3) Non-relatives: $\Delta = (0,0,1)$;
(4) Full sibs: $\Delta = (1/4, 1/2, 1/4)$;
(5) Double first cousins: $\Delta = (1/16, 6/16, 9/16)$;
(6) First cousins or great-grandpare and great-grandchild: $\Delta = (0, 1/4, 3/4)$;
(7) Second cousins: $\Delta = (0, 1/16, 15/16)$;
(8) Grandparent and grandchild, half sibs, or uncle and nephew: $\Delta = (0, 1/2, 1/2)$.

4. A R function for random simulation of relative pair genotype

The function genotype.relatives is written using R soft for random simulation of relative pair genotype as follows.

```
> genotype.relatives=function(n,p,relative=9,delt.2.1.0=c(0,0,0)){
+ vector.BID=matrix(nrow=3,ncol=9)
+ relatives.twins.1=c(1,0,0)
+ relatives.parent.2=c(0,1,0)
+ relatives.not.3=c(0,0,1)
+ relatives.brothers.4=c(1/4,1/2,1/4)
+ relatives.double.first.cousin.5=c(1/16,6/16,9/16)
+ relatives.first.cousin.6=c(0,1/4,3/4)
+ relatives.second.cousin.7=c(0,1/16,15/15)
+ relatives.uncle.nephew.8=c(0,1/2,1/2)
+ vector.BID[,1]=relatives.twins.1
+ vector.BID[,2]=relatives.parent.2
+ vector.BID[,3]=relatives.not.3
+ vector.BID[,4]=relatives.brothers.4
+ vector.BID[,5]=relatives.double.first.cousin.5
+ vector.BID[,6]=relatives.first.cousin.6
+ vector.BID[,7]=relatives.second.cousin.7
+ vector.BID[,8]=relatives.uncle.nephew.8
+ vector.BID[,9]=c(0,0,0)
+ det=vector.BID[,relative]+delt.2.1.0
+ det2=det[1]
+ det1=det[2]
+ det0=det[3]
+ I=diag(rep(1,3))
+ T=matrix(c(p,p/2,0,1-p,1/2,p,0,(1-p)/2,1-p),3)
+ O=matrix(c(rep(p^2,3),rep(2*p*(1-p),3),rep((1-p)^2,3)),3)
+ S=det2*I+det1*T+det0*O
+ C=S
+ C[1,]=p^2*C[1,]
+ C[2,]=2*p*(1-p)*C[2,]
+ C[3,]=(1-p)^2*C[3,]
+ numA=function(a,b)
+ sample(c(2,1,0),size=1,prob = c(a,b,1-a-b))
+ numAA=function(p){
+ d1=numA(p^2,2*p*(1-p))
+ e1=numA(S[3-d1,1],S[3-d1,2])
+ c(d1,e1,C[3-d1,3-e1])}
+ numpair=replicate(n,numAA(p))
+ list(matrix.I=I,matrix.T=T,matrix.O=O,Conditional.probability=S,
+ Joint.probability=C,Joint.genotype.probability=numpair)}
```

For example, we generate 10 genotypes of father-son pairs when $p = 0.7$ as follows.

```
> genotype.relatives(10,0.7,2)

$‘matrix.I‘
[,1] [,2] [,3]
[1,]        1     0     0
[2,]        0     1     0
[3,]        0     0     1
$matrix.T
[,1] [,2] [,3]
[1,] 0.70 0.3 0.00
[2,] 0.35 0.5 0.15
[3,] 0.00 0.7 0.30
$matrix.O
[,1] [,2] [,3]
[1,] 0.49 0.42 0.09
[2,] 0.49 0.42 0.09
[3,] 0.49 0.42 0.09
$Conditional.probability
[,1] [,2] [,3]
[1,] 0.70 0.3 0.00
[2,] 0.35 0.5 0.15
[3,] 0.00 0.7 0.30
$Joint.probability
[,1] [,2] [,3]
[1,] 0.343 0.147 0.000
[2,] 0.147 0.210 0.063
[3,] 0.000 0.063 0.027
$Joint.genotype.probability
[,1] [,2] [,3] [,4] [,5]
[,6] [,7] [,8] [,9] [,10]
[1,] 2.000 2.000 2.000 1.000 1.00 2.000 2.000 2.000 2.000 0.000
[2,] 2.000 2.000 1.000 0.000 1.00 2.000 2.000 1.000 2.000 0.000
[3,] 0.343 0.343 0.147 0.063 0.21 0.343 0.343 0.147 0.343 0.027
```

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 twolocus models using an extension of the li-sacks.method, American journal of human genetics 63 (1998) A297.

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

[7]. W. Weinberg, Uber den nachweis der vererbung beim menschen, Jahreshefte¨ des Vereins fu¨r Vaterl¨andische Naturkunde in Wu¨rttemberg 64 (1908) 368–382.