7 Infinite Populations – Two Loci

7.1 Linkage Disequilibrium

Linkage equilibrium coresponds to the statistical independence of the alleles in two loci of a randomly selected gamete. Unlike the Hardy-Weinberg Equilibrium, linkage equilibrium does not materialize in one generation of random mating. In order to illustrate convergence to linkage equilibrium consider two bi-allelic loci inherited from the same parent. Denote, as before, the alleles of the first locus by A and a and use B and b to denote the alleles of the second locus. Let θ be the recombination fraction between the two loci. Let us examine the probability of the *haplotype* A/B in a given generation, namely the relative frequency in that generation of pairs of loci inherited from the same parent with allele A at the first locus and allele B at the second. Call this probability $p_{\rm AB}$. Under linkage equilibrium this probability equals the product of the marginal probabilities $p_{\rm A}p_{\rm B}$. In general, one may consider the difference $D = p_{\rm AB} - p_{\rm A}p_{\rm B}$ as a measure of linkage disequilibrium in the population. Consider the level of linkage disequilibrium in the next generation following random mating. The haplotype A/B will emerge as a result of one of two possibilities. In the case where recombination does not occur in the parent, the haplotype appears in an offspring if it was inherited from the parent. The probability of inheritance is $p_{\rm AB}$. If recombination in the parent does take place, then the given haplotype will emerge if the allele in the parent at the first locus of one homologous chromosome is A and the allele at the second locus of the other homolog is B. The probability of the first event is $p_{\rm A}$ and the probability of the second event is $p_{\rm B}$. Under random mating, hence Hardy-Weinberg equilibrium, the two loci inherited from the different grandparents are independent. It follows that the probability in the case of recombination is $p_{\rm A}p_{\rm B}$. Combining these arguments, we see that the disequilibrium in the next generation is equal to:

$$\tilde{D} = (1-\theta)p_{\rm AB} + \theta p_{\rm A}p_{\rm B} - p_{\rm A}p_{\rm B} = (1-\theta) \times (p_{\rm AB} - p_{\rm A}p_{\rm B}) = (1-\theta)D \ .$$

By recursion we find that after g generations the level of linkage disequilibrium shrinks to $D_g = (1-\theta)^g \times D_0$. Hence under random mating there is relatively fast convergence to linkage equilibrium; but unlike Hardy-Weinberg equilibrium, linkage equilibrium does not occur in a single generation, even when the two loci are on different chromosomes.

7.2 Estimating Linkage Disequilibrium

Let us assume that we are provided with a sample of unrelated individuals from some target population. A pair of bi-allelic markers are genotyped for each of the individuals in the sample with the goal of determining the distribution of the two-locus haplotypes. In order to clarify the issues involved and to make the presentation more targeted we consider a numerical example as we walk through the details of the discussion. Return to the artificial data presented in the previous chapter, which was saved as an `R` dataframe under the name "`CR`". This dataframe contains genotype information from 3260 individuals collected for three SNPs. In the chapter we tested the association between these SNPs and the disease status and found strong association between `snp1` and the disease and between `snp3` and the disease. In light of these results one may raise a question regarding the relationship between these two markers: are both markers so strongly correlated with each other that each should be considered equivalent to a single marker, or is each marker providing independent information with respect to the association with the disease?

In order to address this question we would like to assess the correlation coefficient r between `snp1` and `snp2`. Here we are motivated by the presumption that a correlation coefficient close to one (in absolute value) is an indication that the first possibility is correct and a correlation coefficient closer to zero supports the second possibility. We are tempted to address the issue of estimating the correlation coefficient from genotypic data by first computing a table of frequencies of haplotype the two SNPs, and then using the estimated distribution in order to compute the correlation coefficient.

Let us initiate the process by taking a second look at the numerical example:

```
> CR <- read.table("CaseRandom.csv",header=TRUE,sep=",")
> table(CR$snp1,CR$snp3)

     C/C T/C T/T
  A/A 172 265 119
  T/A 228 880 492
  T/T  74 440 590
```

The table we formed represents the joint distribution of genotypes in our sample. Each person in the sample appears in one of the table entries. Each person, however, is represented by a pair of haplotypes associated with the pair of copies of the given autosome. In order to make the point, we rewrite

the same table in a different format in Table 1

Table 1: Joint distribution of genotypes for a pair of markers

`snp1`	`snp2`	frequency	haplotype 1	haplotype 2
A/A	C/C	172	A-C	A-C
A/A	T/C	265	A-T	A-C
A/A	T/T	119	A-T	A-T
A/T	C/C	228	A-C	T-C
A/T	T/C	880	A-T and T-C or A-C and T-T?	
A/T	T/T	492	A-T	T-T
T/T	C/C	74	T-C	T-C
T/T	T/C	440	T-T	T-C
T/T	T/T	590	T-T	T-T

Observe that the actual frequency in the sample of haplotypes can be partially inferred from the genotypes. For example, we can infer that each of the 172 subjects with a genotype A/A in `snp2` and a genotype C/C in `snp1` must carry a pair of A-C haplotypes. It is also the case that each of the 265 subjects who are heterozygote at `snp1` but A/A-homozygote at snp2 must carry a single copy of the A-C haplotype (and a single copy of the A-T haplotype). Similarly, it can be inferred that each of the 228 subjects of the 4th row of Table 1 carries a single copy of the A-C haplotype (and a single copy of the T-C haplotype).

However, for the 880 double-heterozygote at the 5th row of the table one cannot determine the haplotype composition, since both the pair (A-T,T-C) and the pair (A-C,T-T) are consistent with the genotype. The other subjects in the sample do not carry the A-C haplotype.

One may conclude, thus, that the frequency of the A-C haplotype may be any number between $2 \times 172 + 265 + 228 = 837$ and $837 + 880 = 1717$, out of a total of $2 \times 3260 = 6520$ haplotypes in the sample.

Denote by $0 \leq \tilde{\vartheta} \leq 1$ the proportion in the sample of double-heterozygote individuals which have the combination (A-C,T-T) of haplotypes. Given the value of $\tilde{\vartheta}$ we can conclude that the frequency of the haplotype A-C in the sample is $837 + \tilde{\vartheta} \times 880$. Likewise, for a given value of $\tilde{\vartheta}$, the frequencies of the other 3 haplotypes in the sample are given by the entries in Table 13.2:

Denote the probabilities of the four haplotypes by p_i, $1 \leq i \leq 4$, according to the four rows of Table 2. Natural estimates of these probabilities as functions of $\tilde{\vartheta}$ are the corresponding relative frequencies, i.e., the entries of the table divided by the total number of 6520. Since we assume that

Table 2: The frequency in the sample of the four haplotypes, given the proportion of (A-C,T-T) double-heterozygotes

Haplotype	Frequency
A-C	$837 + \tilde{\vartheta} \times 880$
A-T	$995 + (1 - \tilde{\vartheta}) \times 880$
T-C	$816 + (1 - \tilde{\vartheta}) \times 880$
T-T	$2112 + \tilde{\vartheta} \times 880$

haplotypes are inherited as a unit (i.e., without recombination), each one can be regarded as a single allele. Assuming they are in Hardy-Weinberg equilibrium, one can express the unknown $\tilde{\vartheta}$ in terms of the allele frequencies. The probability of obtaining a double-heterozygote is $p_1p_4 + p_2p_3$. The probability of a person have the pair of haplotypes (A-C,T-T), given that he or she is double-heterozygote, is:

$$\vartheta = \frac{p_1p_4}{p_1p_4 + p_2p_3} \ . \tag{10}$$

Equating ϑ with $\tilde{\vartheta}$ we get from Table 2 the relation:

$$\vartheta = \frac{(837 + \vartheta\, 880)(2112 + \vartheta\, 880)}{(837 + \vartheta\, 880)(2112 + \vartheta\, 880) + (995 + (1 - \vartheta)\, 880)(816 + (1 - \vartheta)\, 880)} \ , \tag{11}$$

which can be solved in order to obtain a numerical value for ϑ, and thereby numerical values for the haplotype frequencies.

Before providing a more general justification of the proposed procedure let us implement it in the numerical example

```
> N <- c(837,995,816,2112)
> H <- 880
> hetero <- function(th,N,H)
+ {
+     f <- N + c(th,1-th,1-th,th)*H
+     t <- f[1]*f[4]/(f[1]*f[4] + f[2]*f[3])
+     return(th - t)
+ }
> th <- uniroot(hetero,c(0,1),N=N,H=H)$root
> th
[1] 0.7806138
```

The function "`uniroot`" finds in this case the value of "`th`" which solves the given equation. This value can be used to determine the distribution of haplotypes and the value of r:

```
> f <- N + c(th,1-th,1-th,th)*H
> p <- f/sum(f)
> c <- p[1]*p[4]-p[2]*p[3]
> p1 <- p[1]+p[2]
> p2 <- p[1]+p[3]
> r <- c/sqrt(p1*(1-p1)*p2*(1-p2))
> r
[1] 0.3002771
```