6 Infinite Populations – Single Locus

In contrast to the situation in the previous part, where there was a finite, usually very small, mating population, in this part of the course we assume an infinitely large population. Random mating corresponds to the case where the choice of the mating partner is not influenced by the genotypes under consideration. The backcross and intercross are examples of non-random mating. On the other hand, mating in wild populations in general, and in human genetics in particular, is often modeled by random mating.

Random mating enhances the mixing of the genetic material in the population and drives the genotypes toward steady-state conditions. Two important steady states are *Hardy-Weinberg equilibrium* and *linkage equilibrium*. Hardy-Weinberg equilibrium refers to the joint distribution of the two homologous copies of an autosomal allele. In a steady state these two copies are independent and identically distributed. Linkage equilibrium, on the other hand, refers to the joint distribution at two or more loci. In linkage equilibrium the distribution of alleles at different polymorphic loci are independent of each other.

6.1 The Hardy-Weinberg Equilibrium

In order to motivate the idea of Hardy-Weinberg equilibrium consider a single bi-allelic locus with alleles A and a. The genotype of a random individual can be AA, Aa, or aa. Under Hardy-Weinberg equilibrium the frequency of these genotypes are p_{A}^2, $2\,p_{\mathrm{A}}(1-p_{\mathrm{A}})$, and $(1-p_{\mathrm{A}})^2$, respectively, where p_{A} denotes the fraction of allele A in the population and $1-p_{\mathrm{A}}$ the fraction of allele a. These frequencies of genotypes will emerge in offspring of random mating. Indeed, the probability that the father will contribute the allele A to his offspring is p_{A}. The probability of inheriting allele A from the mother is also p_{A}. Random mating corresponds to independence of the two contributions, which taken together produce the indicated binomial probabilities. Note also that when the fractions of the genotypes AA, Aa, and aa are p_{A}^2, $2\,p_{\mathrm{A}}(1-p_{\mathrm{A}})$, and $(1-p_{\mathrm{A}})^2$, the frequencies of the alleles A and a are, respectively, $[2\,p_{\mathrm{A}}^2 + 2\,p_{\mathrm{A}}(1-p_{\mathrm{A}})]/2 = p_{\mathrm{A}}(p_{\mathrm{A}}+1-p_{\mathrm{A}}) = p_{\mathrm{A}}$ and $1-p_{\mathrm{A}}$. Hence the allelic frequencies of the children's generation are the same as in the parents generation, so the Hardy-Weinberg frequencies apply to the genotypes of the grandchildren, etc. This is in contrast to the finite populations we considered earlier in this chapter, where allele frequencies change from one generation to the next and heterozygosity eventually disappears with repeated random mating.

6.2 Derivation of the Hardy-Weinberg Equilibrium

One approach for proving the Hardy-Weinberg Equilibrium relies on the Markovian properties mating types as they evolve through the generations. The central point is that the genotype of the offspring are a function of previous generations only through the genotypes of the parents. The basic tool involves the formation of a mating table, which contains a list of all possible mating types and their frequencies in the population, followed by the examination of the dynamics of the frequencies of the mating types generation by generation. A mating type corresponds to a pairing of two genotypes. Hence, the frequency of the different genotypes in the population at each generation can be read off the table.

In order to simplify the examination let us consider a bi-allelic locus include the frequencies of the genotypes as part of the table:

Mating type	Frequency	A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	p_{11}^2	1	0	0
$A_1A_1 \times A_1A_2$	$p_{11}p_{12}$	0.5	0.5	0
$A_1A_1 \times A_2A_2$	$p_{11}p_{22}$	0	1	0
$A_1A_2 \times A_1A_1$	$p_{12}p_{11}$	0.5	0.5	0
$A_1A_2 \times A_1A_2$	p_{12}^2	0.25	0.5	0.25
$A_1A_2 \times A_2A_2$	$p_{12}p_{22}$	0	0.5	0.5
$A_2A_2 \times A_1A_1$	$p_{22}p_{11}$	0	1	0
$A_2A_2 \times A_1A_2$	$p_{22}p_{12}$	0	0.5	0.5
$A_2A_2 \times A_2A_2$	p_{22}^2	0	0	1

Note that we are using the assumption of random mating in the construction of the table (as well as the assumption that the males and females have identical frequencies of genotypes and that Mendel's first law of equal segregation applies. It is customary to spell out loud the further assumptions of no outside influence in the form of mutations, and migration and the assumption of no survival advantage to any of the genotypes.

From the table we get that

$$\begin{aligned} P(A_1A_1) &= p_{11}^2 + 0.5p_{11}p_{12} + 0.5p_{12}p_{11} + 0.25p_{12}^2 \\ &= (p_{11} + 0.5p_{12})^2 = p_{1.}^2. \end{aligned}$$

Likewise, one can show that $P(A_1A_2) = 2p_{1.}p_{2.}$ and $P(A_2A_2) = p_{2.}^2$, which corresponds to the Hardy-Weinberg assumption. Observe that the fact that the Hardy-Weinberg relation holds did not depend on the initial frequencies of the different genotypes. The derivation just showed can be concluded by

saying that if the listed assumptions hold then the Hardy-Weinberg Equilibrium is reached after one generation.

6.3 Inbreeding in Infinite Populations and Identity by Descent

In finite populations random mating will occasionally produce mating between relatives. Here we consider mating between relatives in infinite populations. Two individuals are related if they have a common ancestor. Siblings and half-siblings have at least one common parent; first cousins have common grandparents, etc. If two individuals have a common ancestor, then it is possible that at a given locus they have inherited the same allele from that ancestor. Such an allele is said to be inherited identical by descent (IBD). The coefficient of relatedness of two individuals is defined to be the probability that at a given locus a randomly selected allele from one of the individuals is identical by descent with a randomly selected allele at the same locus in the other individual. For example, two siblings, whose parents are unrelated, have two common ancestors, their mother and their father. If we select a random allele from one of the siblings, there is probability 1/2 it was inherited from their mother and probability 1/2 it was inherited from their father. In either case, if we select a random allele from the other sibling, there is a 1/2 chance it was inherited from the same parent, and if so, there is then a 1/2 chance it is the same allele. Hence the coefficient of relatedness of the siblings is $2 \times (1/2) \times (1/2) = 1/2$, where the 2 results from having to common ancestors – the two parents. Similar computations show that the coefficient of relatedness of two cosines is 1/8. If two relatives mate, the coefficient of inbreeding of their offspring is by definition the probability of relatedness of the parents, i.e., the probability that at a given locus the two alleles in that offspring are identical by descent.

The following equations modify Hardy-Weinberg equilibrium to accommodate inbreeding. Assume that in a population that is otherwise in Hardy-Weinberg equilibrium, mating occurs between two relatives having a coefficient of relatedness F. Then at a locus with alleles A and a, a child can have a genotype AA because (i) it inherits the A allele from one parent and the same allele IBD from the other parent, which occurs with probability Fp_{A}, or (ii) it inherits the allele A independently (not IBD) from both parents, which happens with probability $(1-F)p_{\mathrm{A}}^2$. Adding these two terms together, we find that the probability of the genotype AA is $p_{\mathrm{AA}} = p_{\mathrm{A}}^2 + Fp_{\mathrm{A}}(1-p_{\mathrm{A}})$. A similar formula holds for the genotype aa. For the genotype Aa, the alleles cannot be inherited IBD, since they are different, so $p_{\mathrm{Aa}} = 2\,p_{\mathrm{A}}(1-p_{\mathrm{A}})(1-F)$.

A consequence of inbreeding is an increase in homozygosity compared to random mating.

Homework Question 6.1. *What is the coefficient of relatedness of two half-siblings, of an aunt and her niece, of two first cousins, of a grandmother and her grandchild? In a child whose parents are first cousins, what is the probability that the child's alleles are IBD at a given autosomal locus?*

Homework Question 6.2. *Assume that the frequency of survival to reproduction age for the three genotypes are w_{11}, w_{12} and w_{22}, respectively. To which values do the frequencies of the genotypes converge? Will the Hardy-Weinberg Equilibrium hold?*

6.4 Statistical tests of Hardy-Weinberg

A statistical test of the Hardy-Weinberg Equilibrium that is based on a sample of unrelated individuals may be obtained as an application of a chi-square test to the frequency table of genotypes. Hence, for example, if a bi-allelic locus is considered, then the frequency table is composed of three cells, one for each genotype. The observed cell counts are compared to the expected counts. The latter are obtained using the observed allele frequencies and the assumed independence. The distance between the two tables is measured with a chi-square statistic, with one degree of freedom in this case.

In order to illustrate the construction of the test of Hardy-Weinberg let us consider the following data from an artificial Case-Control study, which examines 3 SNPs. The data is stored in a text file with fields separated by comas (the `.csv` format). We use the function `read.table` in order to read these data into an R data-frame object:

```
> CR <- read.table("CaseRandom.csv",header=TRUE,sep=",")
> summary(CR)
  group       sex         snp1         snp2         snp3
 CASE: 634   F: 855   A/A: 556   A/A: 701   C/C: 474
 RAND:2626   M:2405   T/A:1600   G/A:1674   T/C:1585
                      T/T:1104   G/G: 885   T/T:1201
```

Examining the data we see that there are five variables listed and 3,260 observations. The primary tests in Case-Control trials involve the examination of the dependence between the phenotype, disease status in this case, and the genotypes:

```
> tab <- table(CR$group,CR$snp1)
> chisq.test(tab)

        Pearson's Chi-squared test

data:  tab
X-squared = 17.1176, df = 2, p-value = 0.0001919
```

However, here we are interested in testing the validity of the Hardy-Weinberg assumption:

```
> tab["RAND",]
 A/A  T/A  T/T
 416 1290  920
>
> p.A <- (tab["RAND",1] + 0.5*tab["RAND",2])/sum(tab["RAND",])
> p.A
[1] 0.4040366
> E <- c(p.A^2,2*p.A*(1-p.A),(1-p.A)^2)*sum(tab["RAND",])
> E
[1]  428.6828 1264.6344  932.6828
> U <- sum((tab["RAND",]-E)^2/E)
> U
[1] 1.056463
> pchisq(U,1,lower.tail=FALSE)
[1] 0.3040234
```

Hence, at least within the given subgroup and for the given SNP, the Hardy-Weinberg assumption is valid.

Note that the number of possible genotypes increases like the square of the number of alleles. Therefore, the power of the test we proposed may be poor if the locus is multi-allelic. Let us consider an alternative test for this case, which is suggested based on the notion of relatedness introduced in the previous subsection.

Specifically, the test is motivated by the departures from equilibrium suggested by the discussion of inbreeding. Suppose alleles at a given locus $i = 1, \ldots, k$ have frequencies p_i. Denote genotypes by *ordered* pairs (i, j), which under Hardy-Weinberg equilibrium have frequency $p_i p_j$. (In reality we do not observe the ordered genotype, but the mathematical notation is simplified if we pretend that we do.) Suppose we have a sample of n genotypes, and X_{ij} is the number of (i, j) genotypes. Consider the inbreeding

model that the probability of genotype (i, j) is given by $p_{ii} = p_i^2 + F p_i(1-p_i)$, while for $i \neq j$, $p_{ij} = p_i p_j (1-F)$. One can write the log-likelihood function, $\ell(F, p_1, \ldots, p_k)$, and show that the efficient score $\partial\ell/\partial F$ evaluated at $F = 0$ is given by $\sum_{i=1}^{k} X_{ii}/p_i - n$. It can be show that the variance of the efficient score when $F = 0$ is $n(k-1)$, which enables standardization of the efficient score.

Note, that we cannot use the efficient score directly as a test statistic because the parameters $p_1, \ldots, p_k$ are usually unknown. However, reasonable estimators of the p_i, when Hardy-Weinberg equilibrium holds, are $\hat{p}_i = (2\,n)^{-1}\big[\sum_{j=1}^{k} X_{ij} + \sum_{j=1}^{k} X_{ji}\big]$. It may be shown, although the argument is still more complicated, that when Hardy-Weinberg equilibrium holds, the statistic $[\sum_{i=1}^{k} X_{ii}/\hat{p}_i - n]/(n(k-1))^{1/2}$ has approximately a standard normal distribution when n is large, and hence can be used as a test of Hardy-Weinberg equilibrium.

Let us illustrate the test in a small simulation. Let consider a polymorphic locus with 5 uniformly distributed distinct alleles and take a sample of 100 subjects:

```
> n <- 100
> p <- rep(1/5,5)
> F <- 1/8
> a1 <- sample(1:5,n,rep=TRUE,prob=p)
> a2 <- sample(1:5,n,rep=TRUE,prob=p)
> IBD <- rbinom(n,1,F)
> a2[IBD==1] <- a1[IBD==1]
> X <- table(a1,a2)
```

Observe that we distinguish between parental alleles, hence we can apply the chi-square test in a straightforward manner:

```
> chisq.test(X)

        Pearson's Chi-squared test

data:  X
X-squared = 20.5693, df = 16, p-value = 0.1957

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(X)
> X
   a2
```

```
a1  1 2 3 4 5
  1 7 2 1 3 2
  2 1 8 7 4 4
  3 4 8 2 4 2
  4 1 5 3 6 2
  5 5 7 4 3 5
```

The warning massage results from the fact that the counts in some of the cells is below five, hence the normal approximation cannot be trusted. As a remedy one may use R's built-in simulation procedure to compute the p-value:

```
> chisq.test(X,sim=TRUE)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  X
X-squared = 20.5693, df = NA, p-value = 0.2104
```

Even with simulations the results obtained are not significant. On the other hand, when we apply the proposed score test we get (borderline) significance:

```
> p.hat <- (colSums(X)+rowSums(X))/(2*n)
> Z <- (sum(diag(X)/p.hat) - n)/sqrt(4*n)
> Z
[1] 2.046907
> 2*(1-pnorm(Z))
[1] 0.04066721
> 1-pchisq(Z^2,1)
[1] 0.04066721
```