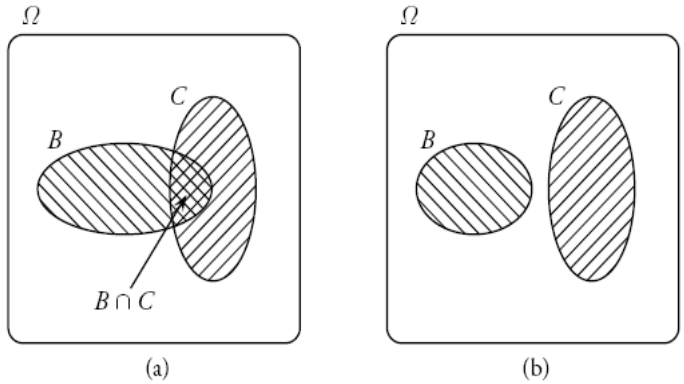
Dominant and recessive alleles

- If a single copy of an allele results in the same phenotype as two copies irrespective of the second allele, the allele is said to be *dominant* over the second allele
- Likewise, an allele which must occur in both copies of the gene to yield the phenotype is termed *recessive*
- Alleles which correspond to mutations which destroy the coding of a protein *tend to be recessive*
- If the phenotype for genotype *i=j* is intermediate between the phenotypes for *i=i* and *j=j*, the alleles *i* and *j* are *codominant*



$B \cup C$ = 'at least one of B and C occur'
$B \cap C$ = 'both B and C occur'
B^* = 'B does not occur'.

General Addition Rule: $P(B \bigcup C) = P(B) + P(C) - P(B \bigcap C)$

Genotype, phenotype, and penetrance

- Because human cells are diploid, there are two alleles at each locus .This pair of alleles is called the individual's *genotype* at that locus
- The *phenotype* is the characteristic (e.g. eye colour) that results from having a specific *genotype*
- Often we require probability models to describe phenotypic expression of genotypes. Probabilities of phenotype conditional upon genotype are called *penetrances*
- In many cases, the same phenotype can result from a variety of different genotypes (called *phenocopies*)
- The same gene may also have several different phenotypic manifestations (called *plieotrophy*)

Allele & genotype frequencies and Heterozygosity

- If the 2 alleles at a locus are the same, the individual is said to be *homozygous* at the locus.
- If they are different, he/she is said to be *heterozygous*
- The *heterozygosity* of a marker is a summary of the allele frequency distribution at a locus in a population.
- Heterozygosity is defined as the probability that two alleles chosen at random from the population are different.

If p_i is the (relative) frequency of the *i*-th allele,

$$heterozygosity = 1 - \sum_i p_i^2$$

Example 6 (Penetrances of a binary disease.) Suppose we have an inheritable monogenic disease, i.e. the susceptibility to the disease depends on the genetoype at one certain locus. Suppose there are two possible alleles A and a at this locus. Usually A denotes the *disease susceptibility allele* and a the *normal allele*, respectively.

$$\begin{aligned} f_0 &= P(\text{'affected'}|(aa)), \\ f_1 &= P(\text{'affected'}|(Aa)), \\ f_2 &= P(\text{'affected'}|(AA)), \end{aligned} \tag{2.4}$$

Two events B and C are *independent* if the occurrence of B does not affect the conditional probability of C and vice versa. In formulas, this is written

$$P(C) = P(C|B) = \frac{P(B \cap C)}{P(B)} \Longleftrightarrow P(B \cap C) = P(B)P(C). \tag{2.5}$$

disease has $f_1 = f_2 = 1$, i.e. one disease allele is sufficient to cause the disease with certainty. However, apart from some genetic traits that are manifest at birth, it is usually the case that $0 < f_0, f_1, f_2 < 1$. The disease is *dominant* if $f_1 = f_2$ and *recessive* if $f_0 = f_1$.

Example 5 (Sibling relative risk.) Given a sib pair, let B and C denote the events that the first and second sibling is affected by a disease respectively. Then

$$K_s = P(C|B)$$

is defined as the *sibling prevalence* of the disease. Whereas the prevalence K_p in (2.1) was the probability that a randomly chosen individual was affected, K_s is the probability of being affected given the extra information that the sibling is affected. For a disease with genetic component(s), we must obviously have $K_s > K_p$. The extent to which the risk increases when the sibling is known to be affected, is quantified by means of the *the relative risk for siblings*,

$$\lambda_s = K_s / K_p. \tag{2.3}$$

The more λ_s exceeds one, the larger is the genetic component of the disease. □

Hardy-Weinberg Equilibrium

In contrast, going from allele frequencies to genotype frequencies requires more assumptions.

HWE Model Assumptions

- infinite population
- discrete generations
- random mating
- no selection
- no migration in or out of population
- no mutation
- equal initial genotype frequencies in the two sexes

Independence & Hardy-Weinberg Proportions

Two events B and C are *independent* if the occurrence of B does not affect the conditional probability of C and vice versa. In formulas, this is written

$$P(C) = P(C|B) = \frac{P(B \cap C)}{P(B)} \Longleftrightarrow P(B \cap C) = P(B)P(C). \tag{2.5}$$

Independence of more than two events can be defined analogously. If $B_1, B_2, \ldots, B_n$ are independent, it follows that

$$P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1) \cdot P(B_2) \cdot \ldots \cdot P(B_n) = \prod_{i=1}^{n} P(B_i).$$

Hardy-Weinberg Equilibrium

In the first generation: $P(A) = u + \frac{1}{2}v$ and $P(a) = w + \frac{1}{2}v$

2nd Generation

Mating Type	Mating Frequency	Expected Progeny
$AA \times AA$	u^2	AA
$AA \times Aa$	$2uv$	$\frac{1}{2}\ AA : \frac{1}{2}\ Aa$
$AA \times aa$	$2uw$	Aa
$Aa \times Aa$	v^2	$\frac{1}{4}\ AA : \frac{1}{2}\ Aa : \frac{1}{4}\ aa$
$Aa \times aa$	$2vw$	$\frac{1}{2}\ Aa : \frac{1}{2}\ aa$
$aa \times aa$	w^2	aa

$*$ Check: $u^2 + 2uv + 2uw + v^2 + 2vw + w^2 = (u + v + w)^2 = 1$

- For the second generation, we have the following genotype frequencies:
  - $p \equiv P(AA) = u^2 + \frac{1}{2}(2uv) + \frac{1}{4}v^2 = \left(u + \frac{1}{2}v\right)^2$
  - $q \equiv P(Aa) = uv + 2uw + \frac{1}{2}v^2 + vw = 2\left(u + \frac{1}{2}v\right)\left(\frac{1}{2}v + w\right)$
  - $r \equiv P(aa) = \frac{1}{4}v^2 + \frac{1}{2}(2vw) + w^2 = \left(w + \frac{1}{2}v\right)^2$
- What are the genotype frequencies in the third generation?

Hardy-Weinberg Equilibrium

When a population is in Hardy-Weinberg equilibrium, the alleles that comprise a genotype can be thought of as having been chosen at random from the alleles in a population. We have the following relationship between genotype frequencies and allele frequencies for a population in Hardy-Weinberg equilibrium:

$$P(AA) = P(A)P(A)$$

$$P(Aa) = 2P(A)P(a)$$

$$P(aa) = P(a)P(a)$$

Hardy-Weinberg Equilibrium

- Consider a locus with two alleles: A and a
- Assume in the first generation the alleles are not in HWE and the genotype frequency distribution is as follows:

1st Generation

Genotype	Frequency
AA	u
Aa	v
aa	w

where $u + v + w = 1$

- From the genotype frequencies, we can easily obtain allele frequencies:

$$P(A) = u + \frac{1}{2}v$$

$$P(a) = w + \frac{1}{2}v$$

Hardy-Weinberg Equilibrium

The frequency of the AA genotype in the third generation is:

$$P(AA) = \left(p + \frac{1}{2}q\right)^2 = \left(\left(u + \frac{1}{2}v\right)^2 + \left(\frac{1}{2}\right)2\left(u + \frac{1}{2}v\right)\left(\frac{1}{2}v + w\right)\right)^2$$

$$= \left(\left(u + \frac{1}{2}v\right)\left[\left(u + \frac{1}{2}v\right) + \left(\frac{1}{2}v + w\right)\right]\right)^2$$

$$= \left(\left(u + \frac{1}{2}v\right)[(u + v + w)]\right)^2$$

$$= \left(\left(u + \frac{1}{2}v\right)1\right)^2 = \left(\left(u + \frac{1}{2}v\right)\right)^2 = p$$

Similarly, $P(Aa) = q$ and $P(aa) = r$ for generation 3

- **Equilibrium** is reached after one generation of random mating under the Hardy-Weinberg assumptions! That is, the genotype frequencies remain the same from generation to generation.

Testing Hardy-Weinberg Equilibrium

- When a locus is not in HWE, then this suggests one or more of the Hardy-Weinberg assumptions is false.
- Departure from HWE has been used to infer the existence of natural selection, argue for existence of assortive (non-random) mating, and infer genotyping errors.
- It is therefore of interest to test whether a population is in HWE at a locus.
- We will discuss the two most popular ways of testing HWE:
  - Chi-Square test
  - Exact test

Hardy-Weinberg Equilibrium

For example, consider a diallelic locus with alleles A and B with frequencies 0.85 and 0.15, respectively. If the locus is in HWE, then the genotype frequencies are:

$$P(AA) = 0.85 * 0.85 = 0.7225$$

$$P(AB) = 0.85 * 0.15 + 0.15 * 0.85 = 0.2550$$

$$P(BB) = 0.15 * 0.15 = 0.0225$$

Example 10 (Heterozygosity of a marker.)

The *heterozygosity* H of a marker is defined as the probability that two independently picked marker alleles are different. It is frequently used for quantifying the degree of polymorphism.

We will apply the law of total probability (2.7), with $B =$ 'the two alleles are of the same type' and $C_i =$ 'allele 1 is of type i'. Then, by the definition of allele frequency $P(C_i) = p_i$. Further, given that C_i has occurred, the event B is the same thing as 'allele 2 is of type i'. Therefore, since the two alleles are picked independently,

$$P(B|C_i) = P(\text{'allele 2 is of type } i'|C_i) = P(\text{'allele 2 is of type } i') = p_i.$$

Finally, we get from (2.7);

$$\begin{aligned} H &= P(B^*) = 1 - P(B) = 1 - \textstyle\sum_{i=1}^{k} P(B|C_i)P(C_i) \\ &= 1 - \textstyle\sum_{i=1}^{k} p_i^2. \end{aligned}$$

The closer to 1 H is, the more polymorphic is the marker.

Theorem 1 (Law of total probability.) *Let $C_1, \ldots, C_k$ be a disjoint decomposition of the sample space*[1]*. Then, for any event B,*

$$P(B) = \sum_{i=1}^{k} P(B|C_i)P(C_i). \tag{2.7}$$

Example 8 (Prevalence under HW equilibrium.) What is the prevalence K_p of a monogenic disease for a population in Hardy-Weinberg equilibrium when the disease allele frequency is $p = 0.02$ and the penetrance parameters are $f_0 = 0.03$, $f_1 = 0.3$ and $f_2 = 0.9$? We apply Theorem 1 with $B =$ 'affected', and C_1, C_2 and C_3 the events that a randomly picked individual has genotype (aa), (Aa) and (AA) respectively

$$\begin{aligned} K_p &= P(B) = P(B|C_1)P(C_1) + P(B|C_2)P(C_2) + P(B|C_3)P(C_3) \\ &= f_0 \cdot (1-p)^2 + f_1 \cdot 2p(1-p) + f_2 \cdot p^2 \\ &= 0.03 \cdot (1-0.02)^2 + 0.3 \cdot 2 \cdot 0.02 \cdot (1-0.02) + 0.9 \cdot 0.02^2 \\ &= 0.0409. \end{aligned} \tag{2.8}$$