SS2857 Probability and Statistics I

Fall 2024

Chapter 2 Summary Exercise: Hardy-Weinberg Equilibrium

Revised 30/09/24

Introduction

As a real-world example of some of the material fro Chapter 2, we are going to consider a problem in population genetics: modelling frequencies of alleles in a population with random mating.

Here is a very brief (and simplified) primer for those of you not familiar with genetics. A gene is a location on a chromosome within an organism that usually contains the code to produce a protein. Alleles for a gene are the possible variants of the DNA sequence that can occur at that location. Different alleles produce different versions of the protein which lead to different physical characteristics. Diploid organisms, including humans, possess two copies of each chromosome and hence two alleles for each gene (barring the genes on the sex chromosomes which may be different, but we will ignore these). The genotype of an individual is determined by the pair of alleles that they inherit from their mother and father. The phenotype is determined by resulting physical characteristic. An allele is said to be dominant if its phenotype is expressed when at least one copy of the allele is present. An allele is recessive if its phenotype is expressed only when both copies of the allele are the same.

A common example is the process for determing eye colour and the alleles for brown versus blue eyes. The allele for brown eyes is dominant to the allele for blue eyes. A person's eyes will be brown (their phenotype) if they have two copies of the brown eye allele or one copy of the brown eye allele and one copy of the blue eye allele (their genotype). The allele for brown eyes is dominant and the allele for blue eyes is recessive. Symbolically, we can let A and a represent the alleles for brown and blue eyes, respectively. The capital letter indicates that the first allele is dominant. The possible genotypes and their correspoding phenotypes are

Genotype	Phenotype
AA	Brown
Aa	Brown
aa	Blue

Notice that the genotypes are unordered. I.e., the genotype doesn't distinguish which allele came from which parent.

Imagine a population in which individuals mate with other individuals in the same generation. I.e., we start with a parent population, mating occurs, and this creates an offspring population. The offspring then go on to mate and produce new offspring, but at no point does an individual mate with an individual from a different generation.

Questions 1

1. Suppose that a population contains only homozygotes (people with genotype *AA* or *aa*). Let n_{AA} and n_{aa} be the numbers of people with genotypes *AA* and *aa* respectively, and let $n = n_{AA} + n_{aa}$ be the total population size.

   Suppose that an offspring is formed from two randomly selected parents. What is the probability that the offspring has each of the genotypes?

Solution: Let G_{AA}, G_{Aa}, and G_{aa} denote the events that the offspring has genotype *AA*, *Aa*, or *aa*. We'll consider *AA* first.

Since there are no heterozygotes (people with genotype *Aa*), the only way that the offspring can have genotype *AA* is if both parents have genotype *AA*. The number of ways this can occur is

$$N(G_{AA}) = \binom{n_{AA}}{2} = \frac{n_{AA}!}{2!(n_{AA}-2)!} = \frac{n_{AA}(n_{AA}-1)}{2}.$$

The total number of combinations of parents is

$$N = \binom{n}{2} = \frac{n!}{2!(n-2)!} \frac{n(n-1)}{2}.$$

Hence, the probability that the inidividual has genotype *AA* is

$$P(G_{AA}) = \frac{n_{AA}(n_{AA}-1)}{n(n-1)}.$$

In exactly the same way,

$$P(G_{aa}) = \frac{n_{aa}(n_{aa}-1)}{n(n-1)}.$$

Then

$$\begin{aligned} P(G_{Aa}) &= 1 - \frac{n_{AA}(n_{AA}-1)}{n(n-1)} - \frac{n_{aa}(n_{aa}-1)}{n(n-1)} \\ &= \frac{n(n-1) - n_{AA}(n_{AA}-1) - n_{aa}(n_{aa}-1)}{n(n-1)}. \end{aligned}$$

2. Suppose that the population is large so that n_{AA}, n_{Aa}, and n_{aa} are all much bigger than 1. Show that

$$P(G_{AA}) \approx p^2, \quad P(G_{Aa)} = 2p(1-p), \quad P(G_{aa)} = (1-p)^2$$

   where p is the proportion of the allele *A* in the parent population.

Solution: The parent population contains $2n_{AA}$ copies of allele *A*, $2n_{aa}$ copies of allele *a*, and $2n$ alleles in total. The proportion of allele *A* is:

$$p = \frac{2n_{AA}}{2n} = \frac{n_{AA}}{n}.$$

Note that the proportion of allele a is

$$\frac{2n_{aa}}{2n} = \frac{n_{aa}}{n} = 1 - p.$$

If n_{AA}, n_{Aa}, and n_{aa} are all much bigger than 1, then

$$n_{AA} - 1 \approx n_{AA}, \quad n_{aa} - 1 \approx n_{aa}, \quad n - 1 \approx n.$$

Substituting these into the probabilities in part a) we get

$$\begin{aligned}
P(G_{AA}) &\approx \frac{n_{AA}^2}{n^2} = p^2 \\
P(G_{aa}) &\approx \frac{n_{aa}^2}{n^2} = (1-p)^2 \\
P(G_{Aa}) &\approx \frac{n^2 - n_{AA}^2 - n_{aa}^2}{n^2} \\
&= \frac{(n_{AA} + n_{aa})^2 - (n_{AA}^2 - n_{aa}^2)}{n^2} \\
&= \frac{2n_{AA}n_{aa}}{n^2} \\
&= 2p(1-p).
\end{aligned}$$

3. Suppose now that the population contains heterozygotes such that n_{AA}, n_{Aa}, n_{aa} be the numbers of people with genotypes AA, Aa, and aa respectively, and let $n = n_{AA} + n_{Aa} + n_{aa}$ be the total population size.

   Suppose that an offspring is formed from two randomly selected parents. Show that the same result occurs.

Solution: First note that the proportion of allele A in the population is

$$p = \frac{2n_{AA} + n_{Aa}}{2n}.$$

Consider genotype AA. There are four possible ways that the offspring can have this genotype:

1. Both parents have genotype AA.
2. The first parent has genotype Aa and contributes allele A and the second has genotype AA.
3. The first parent has genotype AA and the second parent has genotype Aa and contributes allele A.
4. Both parents have genotype Aa and contribute allele A.

Note that these events are mutually exclusive, so we can obtain the total probability by summing their individual probabilities. The probabilities of these four events are:

1. $\frac{n_{AA}\times(n_{AA}-1)}{n(n-1)}$

2. $\frac{n_{AA}\times n_{Aa}/2}{n(n-1)}$

3. $\frac{n_{Aa}/2\times n_{AA}}{n(n-1)}$

4. $\frac{n_{Aa}/2\times(n_{Aa}-1)/2}{n(n-1)}$

Summing these gives

$$\begin{aligned}P(G_{AA}) &= \frac{n_{AA}\times(n_{AA}-1)}{n(n-1)} + 2\frac{n_{AA}\times n_{Aa}/2}{n(n-1)} + \frac{n_{Aa}/2\times(n_{Aa}-1)/2}{n(n-1)} \\ &= \frac{n_{AA}\times(n_{AA}-1)+n_{AA}n_{Aa}+n_{Aa}(n_{Aa}-1)/4}{n(n-1)}.\end{aligned}$$

Once again, we'll assume that the population size is large so that $n_{AA}-1\approx n_{AA}$ and $n_{Aa}-1\approx n_{Aa}$. Then

$$\begin{aligned}P(G_{AA}) &\approx \frac{n_{AA}^2+n_{AA}n_{Aa}+n_{Aa}^2/4}{n(n-1)} \\ &= \frac{(n_{AA}+n_{Aa}/2)^2}{n^2} \\ &= p^2.\end{aligned}$$

In exactly the same way

$$P(G_{aa})\approx(1-p)^2$$

and then

$$P(G_{Aa})\approx 1-p^2-(1-p)^2 = 2p(1-p).$$

Hardy-Weinberg Equilibrium

The results above describe model the frequency of alleles in a population undergoing random mating. Random mating occurs under two conditions:

1. the two alleles an offspring inherits from each parent are independent, and

2. the probability that each allele takes a specific form is equal to the proportion of that allele in the parent population. I.e., the probability that either of an offspring's alleles is A is p and the probability that either is a is $1-p$.

Neither of these conditions can ever be exactly true. The alleles an offspring inherits from its parents are not ever completely independent because selecting one allele from the parent population changes the proportions of the alleles that remain. However, the conditions will be approximately true in a large population. If there are many individuals then choosing one allele from one parent changes the proportions of the remaining alleles only very slightly. In this case, the proportion of individuals with the possible genotypes remains the same from one generation to the next. This is know as Hardy-Weinberg equilibrium.

This is important because the assumption of Hardy-Weinberg allows us to answer other questions about the genetic composition of the population and how mating is occurring without having to genotype every indivdiual (an expensive proposition). Here are some examples:

Questions 2

4. Suppose that a population in Hardy-Weinberg equilibrium contains n_B people with brown eyes and $n_b = n - n_B$ people with blues. What is the probability that a randomly selected person has each of the possible genotypes?

Solution: If the population is in Hardy-Weinberg equilibrium then we know that

$$P(G_{aa}) = (1-p)^2 = \frac{n_b}{n}.$$

This implies that

$$p = 1 - \sqrt{\frac{n_b}{n}}.$$

Then

$$P(G_{AA}) = \left(1 - \sqrt{\frac{n_b}{n}}\right)^2$$

and

$$P(G_{Aa}) = \sqrt{\frac{n_b}{n}}\left(1 - \sqrt{\frac{n_b}{n}}\right).$$

5. Suppose that a population in Hardy-Weinberg equilibrium contains n_B people with brown eyes and $n_b = n - n_B$ people with blues. What is the probability that a randomly selected person has each of the possible genotypes given that they have brown eyes?

Solution: The event that the individual has brown eyes is $G_{AA} \cup G_{Aa}$. Then

$$\begin{aligned} P(G_{AA}|G_{AA} \cup G_{Aa}) &= \frac{P(G_{AA} \cap (G_{AA} \cup G_{Aa}))}{P(G_{AA} \cup G_{Aa})} \\ &= \frac{P(G_{AA})}{P(G_{AA}) + P(G_{Aa})} \end{aligned}$$

since G_{AA} and G_{Aa} are disjoint. Hence

$$\begin{aligned} P(G_{AA}|G_{AA} \cup G_{Aa}) &= \frac{\left(1 - \sqrt{\frac{n_b}{n}}\right)^2}{\left(1 - \sqrt{\frac{n_b}{n}}\right)^2 + \sqrt{\frac{n_b}{n}}\left(1 - \sqrt{\frac{n_b}{n}}\right)} \\ &= \frac{\left(1 - \sqrt{\frac{n_b}{n}}\right)}{\left(1 - \sqrt{\frac{n_b}{n}}\right) + \sqrt{\frac{n_b}{n}}} \\ &= 1 - \sqrt{\frac{n_b}{n}}. \end{aligned}$$

Similarly

$$\begin{aligned} P(G_{Aa}|G_{AA} \cup G_{Aa}) &= \frac{\sqrt{\frac{n_b}{n}}\left(1 - \sqrt{\frac{n_b}{n}}\right)}{\left(1 - \sqrt{\frac{n_b}{n}}\right)^2 + \sqrt{\frac{n_b}{n}}\left(1 - \sqrt{\frac{n_b}{n}}\right)} \\ &= \frac{\frac{n_b}{n}}{\left(1 - \sqrt{\frac{n_b}{n}}\right) + \sqrt{\frac{n_b}{n}}} \\ &= \sqrt{\frac{n_b}{n}}. \end{aligned}$$

This question could also be answered using the complement rule:

$$P(G_{Aa}|G_{AA}\cup G_{Aa}) = 1 - P(G_{AA}|G_{AA}\cup G_{Aa}).$$

Finally, for completeness:

$$P(G_{aa}|G_{AA}\cup G_{Aa}) = 0.$$