BST227 - Homework 1 Solution

Introduction

Pedigrees have long been the backbone of statistical genetics to infer modes of inheritance for Mendelian disorders, and we'll show later in the course that they can also be quite useful for understanding more complex traits. The first part of this assignment will focus on understanding basic pedigree analyses.

Secondly, we will examine Hardy-Weinberg equilibrium, which has historically been a means to understand alleles in a population and more recently been applied to high-throughput genotyping data for quality control.

Problem 1

Use the pedigree shown in Figure 1 to answer the questions associated with problem 1.

(A) What is the likely form of inheritance for the trait shown on this pedigree?

It is most likely autosomal recessive. There is no gender difference among the affected in the pedigree, so it is likely to be an autosomal mode of inheritance as opposed to sex-linked inheritance. Affected individuals generally do not have a recessive parent, a clear sign for a recessive mode of inheritance.

(B) List the genotypes of individuals 1-5. If multiple genotypes per person are possible, list each.

Here we are denoting 'a' as the causal disease allele.

1) AA, Aa
2) AA, Aa
3) aa
4) Aa
5) AA, Aa

(C) What is the probability that individual 5 will be a carrier of a causal allele?

Individual 5's father is either a carrier or homozygous dominant (the father is unaffected). Moreover, her mother is also a carrier (because the mother is an unaffected individual who _must_have inherited a recessive allele from her affected father). Given that individual 5 is unaffected, she cannot be homozygous recessive. Since this is most likely a pedigree for a rare Mendelian disorder, we can use the rare disease assumption to assume that the father of 5 who marries into the family is not a carrier, AA. If he is not a carrier, then the probability that individual 5 is a carrier with parents AA x Aa is then simply 0.5. If one does not use the rare disease assumption then more generally, one can work with the assumption that p is the frequency of the disease allele in the population and give an expression for individual 5 being a carrier in terms of p.

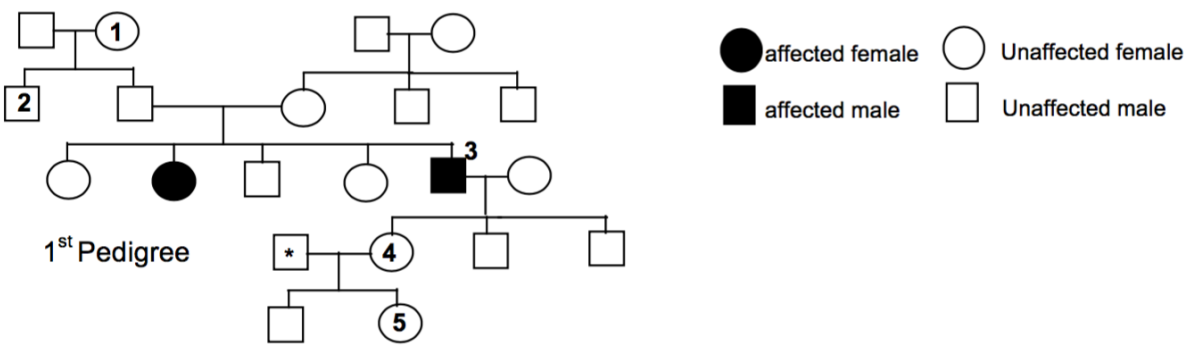

Figure 1: Pedigree for problem 1

Problem 2

Use the pedigree shown in Figure 2 to answer the questions associated with problem 2.

(A) What is the likely form of inheritance for the trait shown on this pedigree?

It is most likely autosomal dominant inheritance. There is no gender difference among the affected individuals in the pedigree, so it is likely to be an autosomal mode of inheritance rather than sex-linked. Affected individuals have an affected parent as well, a clear sign for a dominant mode of inheritance.

(B) List the genotypes of individuals 1-5. If multiple genotypes per person are possible, list each.

Here, we are denoting 'A' as the causal disease allele.

1) aa
2) Aa
3) aa
4) Aa
5) aa

(C) What is the probability that individual 5 will be a carrier of a causal allele?

The probability that individual 5 is a carrier is 0 because individual 5 is unaffected for an autosomal dominant trait. Assuming complete penetrance of the disease allele, it is impossible for her to be a carrier of the disease allele and not express the diseased phenotype.

(D) If individuals 2 and 3 were to have many more children, what are all of the possible genotypes of the offspring and in what proportion would they occur in?

Based on the genotypes inferred in (B), and random assortment in Mendels' Laws, genotypes Aa and aa will each occur with a probability of 0.5.

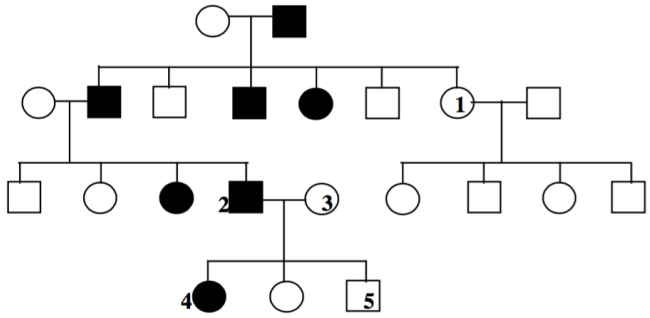

Figure 2: Pedigree for problem 2

Problem 3

What genetic factors must be occurring for a Hardy-Weinberg equilibrium to exist?

The three main genetic factors that must hold for Hardy Weinberg equilibrium to exist are random mating in population, no mutations in the population and no selective pressure on the alleles. At the population genetic level, one also needs assumption of large population size, discreteness of generations, and no migration. However, it is to be noted that all HWE requires is for these factors to hold at the most recent generation as HWE can be established in a single generation.

Problem 4

Suppose that an allele `t` occurs with a frequency of 0.9 in a population of Harvard students. Give the frequency of genotypes `TT`, `Tt`, and `tt`.

$p = 0.1$

$q = 0.9$

Frequency of `TT` $= p^2 = 0.1^2 = 0.01$

Frequency of `Tt` $= 2pq = (2)(0.1)0.9 = 0.18$

frequency of `tt` $= q^2 = 0.9^2 = 0.81$

Problem 5

Suppose that you get genotyping data for a particular variant in a cohort where 4000 individuals are genotypes for `G/G`, 1300 are genotyped for `A/A`, and 3700 are genotyped for `A/G`. Compute the Chi-squared statistic and p-value associated with whether this genotyped variant

follows HWE. Interpret your summary statistic in the context of whether or not you believe that the variant was accurately genotyped.

Frequency of `G` allele: $2 * 4000 + 3700 = 11700$.

Frequency of the `A` allele: $2 * 1300 + 3700 = 6300$.

$p = 11700/(11700 + 6300) = 0.65$

$q = 11700/(11700 + 6300) = 0.35$

Assuming HWE, the expected distribution of the genotypes is:

Expected `G/G`: $np^2 = 9000 * 0.65^2 = 3802.5$
Expected `A/A`: $nq^2 = 9000 * 0.35^2 = 1102.5$
Expected `A/G`: $npq = 9000 * 0.35 * 0.65 = 4095$

Using `R`, we can compute the χ^2 statistic

```
obs <- c(4000, 1300, 3700)
exp <- c(3802.5, 1102.5, 4095)
chisq_stat <- sum((obs-exp)^2/exp)
chisq_stat
```

```
## [1] 83.73922
```

```
pchisq(chisq_stat, 1, lower.tail=FALSE)
```

```
## [1] 5.645418e-20
```

The χ^2 statistic is close to zero, so the distribution of genotypes that we observed deviates significantly from the expected distribution assuming HWE.

The fact that minor allele frequency (MAF) is 0.35 suggests that the SNP is unlikely to be under strong selective pressure. Thus, the large χ^2 value is more likely the result of a technical error in genotyping the variant than an indicator of a departure from HWE for genetic or selective reasons.

Problem 6

Suppose that you are studying a genetic disorder in homozygous recessives that causes death during the teenage years. If 10 in 10,000 newborn babies have the disease, what are the expected frequencies of the three genotypes in newborns, assuming the population is at Hardy-Weinberg equilibrium?

$q^2 = 10/10000 = 0.001$

$q = \sqrt{0.001} = 0.032$

$p = 1 - q = 1 - 0.032 = 0.968$

Proportion of diseased: $q^2 = 0.001$

Proportion of healthy carriers: $2pq = 2 * 0.968 * 0.032 = 0.061$

Proportion of healthy non-carriers: $p^2 = (0.968)^2 = 0.938$

Why is this assumption of Hardy-Weinberg equilibrium not strictly correct for this application?

As the disorder causes death during teenage years, the "no selection" assumption for HWE is violated. This would affect the observed frequencies in two ways. First, if we were sampling from adult individuals, the

recessive genotype would be under-sampled. Secondly, since affected individuals often pass away before they reach reproductive age, there is a selective pressure against the disease allele since affected parents are unable to pass it on to subsequent generations.