BST227 Lab 1

Important Concepts

- Pedigree analysis and interpretation
- HWE and its modern applications
- Computing goodness-of-fit statistics
- Installing `R`, `RStudio`, and `R markdown`
- Using `knitr` to generate code-embedded documents, including `.pdf` files

Pedigree analysis

Modes of Inheritance

- Inheritance patterns describe how a disease is transmitted in families.
- In general, inheritance patterns for single gene disorders are classified bsaed on whether they are autosomal or X-linked and whether they have a dominant or recessive pattern of inheritance.

Autosomal Dominant Inheritance

- Only one copy of a disease allele is necessary for an individual to be susceptible to expressing the phenotype.
- With each pregnancy, there is a one in two chance the offspring will inherit the disease allele.
- Unless a new mutation has occurred, all affected individuals will have at least one parent who carries the disease allele.
- Across a population, the proportion of affected males should equal the proportion of affected females.

Autosomal Recessive

- Two copies of a disease allele are required for an individual to be susceptible to expressing the phenotype.
- Typically, the parents of an affected individual are not affected, but are gene carriers.
- With each prenancy of carrier parents, there is a one in four chance the offspring will inherit two copies of the disease allele and therefore have the phenotype.
- As with autosomal dominant inheritance, the proportion of affected males should be equal to the proportion of affected females in a population.
- More frequently observed in individuals who are descendants of the same ancestors.

X-Linked Dominant Inheritance

- Only one copy of a diease allele on the X chromosome is required for an individual to be susceptible.
- Both males and females can be affected, although males may be more severly affected because they only carry one copy of genes found on the X chromosome.
- If a male is affected, how will this affects his daughters? His sons?

X-Linked Recessive Inheritance

- Two copies of a disease allele on the X chromosome are required for a female to be affected, whereas only one copy of the disease allele on the X chromosome is required for a male to be affected.
- Affected males are related through carrier females.
- If a woman has two copies of the disease allele, how will her sons be affected? Her daughters?

Pedigree Example

Below is an example of a pedigree. Try and infer the mode of inheritance; specifically, is the trait sex-linked or autosomal-linked and is it dominant or recessive.

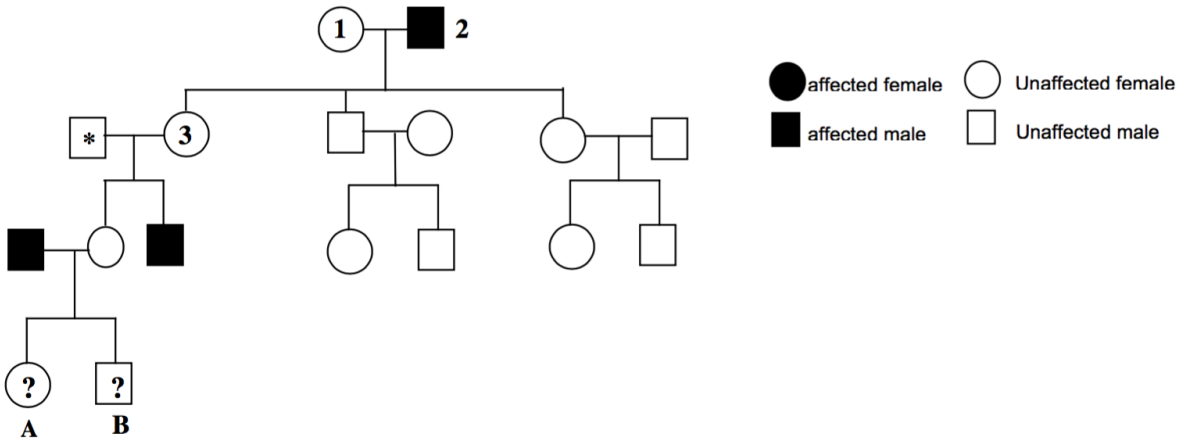

Figure 1:

Questions

1) What is the most likely mode of inheritance of this disease? Choose from: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive.

2) List all possible genotypes of for individuals 1,2,3 in the tree.

Individual 1:

Individual 2:

Individual 3:

3) What is the probability of Individual A being affected?

4) What is the probability of Individual B being affected?

Hardy-Weinberg Equilibrium

- In 1908 Godfrey Hardy and Wilhelm Weinberg independently derived a formula relating allele frequency in parents to genotype frequency in offspring.
- Many assumptions are required for the formula to hold:
  - Random mating
  - No inbreeding
  - Infinite population size
  - Discrete generations
  - Equal allele frequencies in males and females
  - No mutation, migration, or selection (meaning that certain alleles do not confer a selective advantage or disadvantage in reproduction)
- Even though none of these assumptions is likely to hold exactly in any population, the Hardy-Weinberg principle often provides a good approximation for population genotype frequencies.
- Let p be the frequency of the A allele in a population satisfying the assumptions above. A population is said to be in *Hardy-Weinberg Equilibrium (HWE)* if the genotypes in the entire population satisfy the following three conditions:
  - P(AA genotype) = p^2
  - P(Aa genotype) = $2pq$
  - P(aa genotype) = q^2
- Can be useful for detecting:
  - Population substructure
  - Novel mutations
  - Migration status
  - Selection
  - Genotyping errors
- Applications
  - Either a population is assumed to be in HWE, in which case the genotype frequencies can be calculated
  - Or, if the genotype frequencies of all three genotypes are known, they can tested for deviations that are statistically significant.

Example

The establishment of the genetics of the ABO blood group system was one of the first breakthroughs in Mendelian genetics. The locus corresponding to the ABO blood group has three alleles, A, B and O and is located on chromosome 9q34. The alleles A and B are dominant to O. This leads to the following genotypes and phenotypes:

Genotype	Phenotype
AA/AO	A
BB/BO	B
AB	AB
OO	OO

Mendel's first law allows us to quantify the types of gametes an individual can produce. For example, an individual with type AB produces gametes A and B with equal probability $\frac{1}{2}$.

From a sample of 21,104 individuals from the city of Berlin, allele frequencies have been estimated to be $p_A = 0.2877$, $p_B = 0.1065$ and $p_O = 0.6057$. If an individual has blood type B, what gametes can be produced and with what frequency?

Chi-Square goodness of fit test

In classical genetics, Hardy-Weinberg equilibrium was useful for determing allele frequencies and putative markers for selection. Another use in modern genetics / genotyping is the identification of variants that are poorly genotyped when their distributions significantly deviate from the HWE assumption. The simple derivation of a test statistic for the goodness-of-fit

Genotype	Observed	Expected
AA	n_1	np^2
Aa	n_2	$2np(1-p)$
aa	n_3	$n(1-p)^2$
Total	n	n

In this setting:

$$Q = \sum_{geno} \frac{(observed - expected)^2}{expected}$$

$$Q \sim \chi^2_{(1)}$$

The sampling distribution of the test statistic under the null hypothesis is approximately a χ^2 distribution with 1 degree of freedom.

A rule of thumb for χ^2 tests: the expected count should be at least 5 in every cell. If allele frequencies are low, and/or sample size is small, and/or there are many alleles at a locus, this may be a problem and the chi-square distribution may no longer be appropriate. Instead, the Fisher's exact test should be used.

R / R Studio

- Download and install `R`, a free computing environment from the official R Project here.

- One useful way of interfacing with the `R` computing environment is using RStudio. The assignments and labs for this course will rely on `RStudio`, which simply provides an attractive interface for the `R` language.
- You may have `R` installed previously on your machine or may have multiple versions. Ensure that you have `R version > 3.4.0` installed

knitr

The `knitr` package enables dynamic report generation with `R`. Practically, this means that you can create assignment solutions that imbed both code and figures that respond to the question prompt. You can install knitr by typing the following into your `R` console:

```
install.packages("knitr")
```

See this document for a quick overview of how to get started using knitr: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf