Module 5 - Likelihood Over Pedigrees

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Outline

Likelihood Over Pedigrees (Expanding Chapter 4.1 Preliminaries)

- Joint distribution of phenotype Y and genotype G in nuclear families
- Marginal distribution of parents G
- Conditional distribution of offsprings G given parents G
- Conditional probability (penetrance) of Y given G
- Discussion of joint distribution of Y and G in general pedigrees

What's Value of Learning This? e.g. Genetic Counselling

Cystic Fibrosis (CF) is a recessive disease.

- If both parents are carriers, what's the chance that their first born will have CF?
- If both parents are carriers, what's the chance that both their kids will have CF?
- If both parents do not have CF, what's the chance that their first born will have CF?

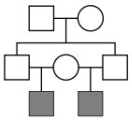
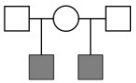
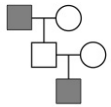
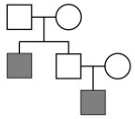
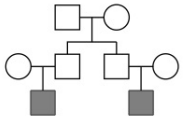
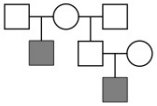
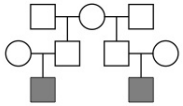
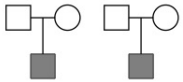
Likelihood/Probability Over Pedigree Data

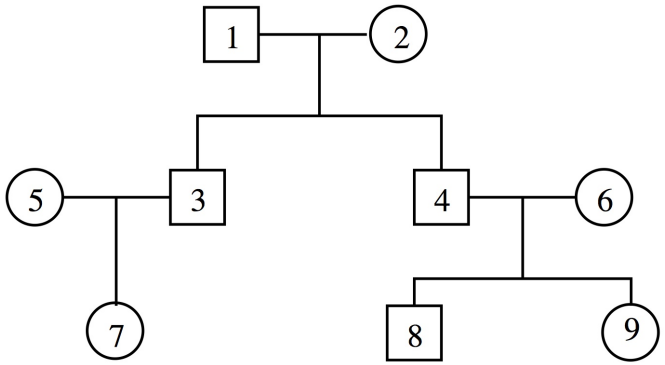
- *The* **approach used to construct a likelihood for pedigree data given in Section 4.1 serves as a basis** *for other analyses in linkage and association discussed in later chapters.*

- Simple nuclear family/pedigree only:
  - Two parents, two offsprings.
  - Female: circle
  - Male: square

- More complex pedigrees discussed in more advanced stat gene course.

- Some pedigree can be really COMPLEX, e.g. The Hutterites data (Dr. Carole Ober, University of Chicago): single one 13-generation pedigree with 1623 individuals (descendants of 64 Hutterite ancestors; quite a bit inbreeding).

Pedigree	Relationship
	MZ-twin
	parent-offspring
	full-sib
	half-sib+first-cousin
	half-sib
	grandparent-grandchild
	avuncular
	first-cousin
	half-avuncular
	half-first-cousin
	unrelated



Joint Distribution of Y and G - Notations

- d and D: two alleles of a biallelic marker.

- $G = \{dd, dD, DD\}$: the three genotypes.
- $\{0, 1, 2\}$: alternative (and conventional) way to code the genotypes; counting the number of copies of the D allele.
- p: the allele frequency of allele D.

- X_1, X_2: genotype variables for siblings 1 and 2, $X_i \in \{0, 1, 2\}$.
  Y_1, Y_2: phenotype variables for siblings 1 and 2.
- P_1, P_2: genotype variables for parents 1 and 2, $P_i \in \{0, 1, 2\}$.

- $f(.)$: probability density function, e.g. $P(X_1 = x_1) = f(x_1)$.

- Convention: capital letters (e.g. X_1 and X_2) denote random variables, and lower case letters (e.g. x_1 and x_2) denote the particular values of the random variables.

Joint Distribution of Y and G I

Several systematic steps involved in the joint distribution:

$$f(y_1, y_2, x_1, x_2, g_1, g_2) = P(Y_1 = y_1, Y_2 = y_2, X_1 = x_1, X_2 = x_2, P_1 = g_1, P_2 = g_2),$$

$$\text{e.g. } P(Y_1 = 1, Y_2 = 1, X_1 = dD, X_2 = DD, P_1 = dD, P_2 = DD) = ?$$

- **Parents** G assuming random mating:

$$f(g_1, g_2) = P(P_1 = g_1, P_2 = g_2) = P(P_1 = g_1)P(P_2 = g_2) = f(g_1)f(g_2).$$

- If we assume HWE then e.g.

$$P(P_1 = 1, P_2 = 2) = P(P_1 = dD, P_2 = DD) = P(P_1 = dD)P(P_2 = DD)$$
$$= 2p(1-p)p^2 = 2p^3(1-p).$$

- Offsprings G: **Note that there is NO independence here!**

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) \neq P(X_1 = x_1)P(X_2 = x_2).$$

Joint Distribution of Y and G II

- **Offsprings G conditional on parents G**, assuming independent Mendelian segregation between offsprings.

$$f(x_1, x_2|g_1, g_2) = P(X_1 = x_1, X_2 = x_2|P_1 = g_1, P_2 = g_2)$$

$$= P(X_1 = x_1|P_1 = g_1, P_2 = g_2)P(X_2 = x_2|P_1 = g_1, P_2 = g_2)$$

$$= f(x_1|g_1, g_2)f(x_2|g_1, g_2).$$

- If we assume Mendelian first law of segregation then e.g.

$$P(X_1 = 0, X_2 = 2|P_1 = 1, P_2 = 2) = P(X_1 = dd, X_2 = DD|P_1 = dD, P_2 = DD)$$

$$= P(X_1 = dd|P_1 = dD, P_2 = DD)P(X_2 = DD|P_1 = dD, P_2 = DD) = 0!$$

$$P(X_1 = 1, X_2 = 2|P_1 = 1, P_2 = 2) = P(X_1 = dD, X_2 = DD|P_1 = dD, P_2 = DD)$$

$$= P(X_1 = dD|P_1 = dD, P_2 = DD)P(X_2 = DD|P_1 = dD, P_2 = DD) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

Joint Distribution of Y and G III

- **Joint offsprings and parents** G (Textbook Equation 4.1),

$$f(x_1, x_2, g_1, g_2) = P(X_1 = x_1, X_2 = x_2, P_1 = g_1, P_2 = g_2)$$

$$= P(X_1 = x_1, X_2 = x_2 | P_1 = g_1, P_2 = g_2) P(P_1 = g_1, P_2 = g_2)$$

$$= f(x_1 | g_1, g_2) f(x_2 | g_1, g_2) f(g_1) f(g_2).$$

- e.g.

$$P(X_1 = 1, X_2 = 2, P_1 = 1, P_2 = 2)$$

$$= P(X_1 = 1, X_2 = 2 | P_1 = 1, P_2 = 2) P(P_1 = 1, P_2 = 2) = \frac{1}{4} 2p^3(1-p).$$

- Again, not all combinations are possible, e.g.

$$P(X_1 = dd, X_2 = DD, P_1 = dD, P_2 = DD) = 0.$$

Joint Distribution of Y and G IV

- There are two components in the joint probability formulation:

$$P(X_1 = x_1, X_2 = x_2, P_1 = g_1, P_2 = g_2)$$

$$= \underline{P(X_1 = x_1, X_2 = x_2 | P_1 = g_1, P_2 = g_2)}\ \underline{P(P_1 = g_1, P_2 = g_2)}$$

  - Parents generation, $P(P_1, P_2)$, needs concepts and models from population genetics (e.g. population allele frequency, HWE).
  - Offsprings generation, $P(X_1, X_2)$, needs concepts and models from DNA segregation transmission between generations, $P(X_1, X_2 | P_1, P_2)$ (e.g. Mendel's first law of segregation).
  - Use of Bayes' rule for the conditional probability.

Joint Distribution of Y and G V

- How do we get the marginal distribution of offsprings G?

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$
$$= \sum_{g_1, g_2} P(X_1 = x_1, X_2 = x_2, P_1 = g_1, P_2 = g_2)$$
$$= \sum_{g_1, g_2} f(x_1|g_1, g_2) f(x_2|g_1, g_2) f(g_1) f(g_2),$$

where $g_1, g_2 \in \{dd, dD, DD\}$ or $\{0, 1, 2\}$.

  - Calculations can be done by hand: a bit tedious, essentially going through the rows in Textbook Table 2.1, or
  - Write a programming script: more efficient and adaptive and less prone to errors.

- A little exercise: $P(X_1 = 2, X_2 = 1) = P(X_1 = DD, X_2 = Dd) =$??? (Check Textbook Page 50.)

Joint Distribution of Y and G VI

- **Adding the phenotype Y component** (Textbook Equation (4.3)).

$$f(y_1, y_2, x_1, x_2, g_1, g_2) = f(y_1, y_2|x_1, x_2, g_1, g_2)f(x_1, x_2, g_1, g_2)$$

$$= f(y_1|x_1)f(y_2|x_2)f(x_1|g_1, g_2)f(x_2|g_1, g_2)f(g_1)f(g_2).$$

- **Important assumptions needed for**

$$f(y_1, y_2|x_1, x_2, g_1, g_2) = f(y_1, y_2|x_1, x_2) = f(y_1|x_1)f(y_2|x_2).$$

  - A person's phenotype Y depends only on the genotype G of one single DSL of that individual. Reasonable only for simple Mendelian disorders.
  - If there are environmental Es factors, Es tend to be correlated within family:

$$f(e_1, e_2) \neq f(e_1)f(e_2),$$

    and we need more complex statistical models.

  - Note that only if $E \perp X$, $f(e_1, e_2|x_1, x_2) = f(e_1, e_2)$.

Joint Distribution of Y and G VII

$$
\begin{aligned}
f(y_1, y_2|x_1, x_2) &= \sum_{e_1, e_2} f(y_1, y_2, e_1, e_2|x_1, x_2) \\
&= \sum_{e_1, e_2} f(y_1, y_2|e_1, e_2, x_1, x_2) f(e_1, e_2|x_1, x_2) \\
&= \sum_{e_1, e_2} f(y_1|e_1, x_1) f(y_2|e_2, x_2) f(e_1, e2) \\
&\neq \sum_{e_1} f(y_1|e_1, x_1) f(e_1) \sum_{e_2} f(y_2|e_2, x_2) f(e_2)
\end{aligned}
$$

- If there are more than one disease susceptibility locus Gs, then Gs can be also correlated.

Back to the Genetic Counselling Example

Cystic Fibrosis (CF) is a recessive disease.

So the penetrances linking genotype G and phenotype Y are:

$$f_0 = P(Y = 1|g = dd) = 0, f_1 = P(Y = 1|g = dD) = 0, f_2 = P(Y = 1|g = DD) = 1.$$

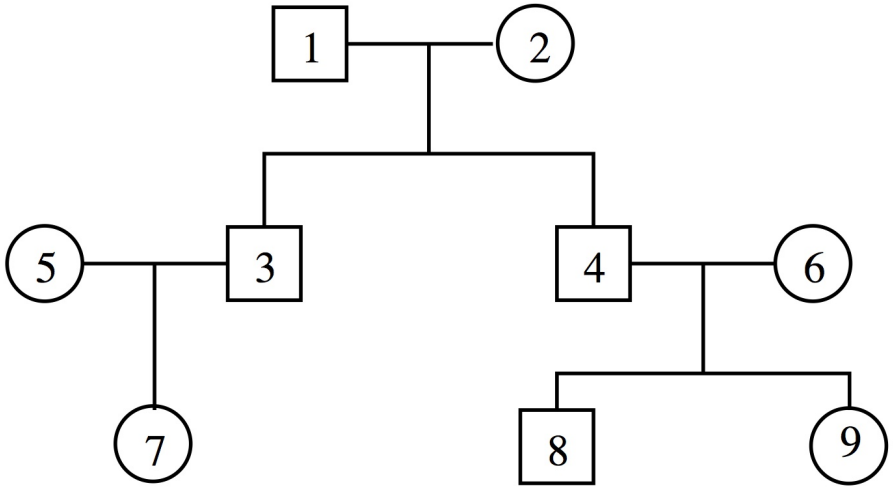
- If both parents are carriers, what's the chance that their first born will have CF?

  A carriers means the genotype is Dd, and using the notations as before: Y for phenotype of an offspring, X for the genotype of the offspring and g_1 and g_2 for the genotypes of the two parents, then

$$\begin{aligned} & P(Y_1 = 1|g_1 = Dd, g_2 = Dd) \\ = \; & P(Y_1 = 1, X_1 = DD|g_1 = Dd, g_2 = Dd) + 0 + 0 \\ = \; & P(Y_1 = 1|X_1 = DD)P(X_1 = DD|g_1 = Dd, g_2 = Dd) \\ = \; & \frac{1}{4} \end{aligned}$$

- If both parents are carriers, what's the chance that both their kids will have CF?
- If both parents do not have CF, what's the chance that their first born will have CF?

Joint Distribution of Y and G - Beyond Siblings I



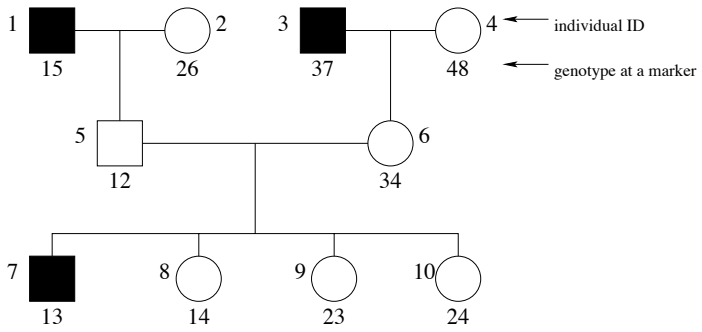
- **Likelihood calculation over general pedigrees** is a more advanced topic, requiring e.g.
  - Elston-Stewart Peeling algorithm.
  - Lander-Green Hidden Markov Model (HMM) algorithm.

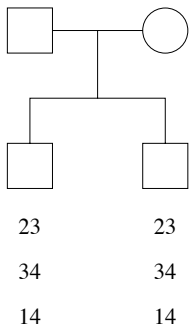
Joint Distribution of Y and G - Beyond Siblings II

- We will touch on the basic ideas of these two algorithms in the linkage analysis (Parametric and nonparametric linkage analysis: a unified multipoint approach.), but not the mathematical details.
  - Need the concept of Markov and Hidden Markov Models (HMM). A tutorial on Hidden Markov Models and selected applications in speech recognition.
  - Need understanding of multi-locus inheritance model for jointly analyzing multiple genetic markers(later).

Joint Distribution of Y and G - Beyond Siblings III

- e.g. more advanced learning will allow us to
  - perform a two-point linkage analysis using the ES peeling algorithm. That is, calculate the likelihood for θ, the recombination fraction between the marker locus, and the unknown gene locus (left figure) and
  - calculate the posterior probability distribution of genetic material shared by the sib pair at each marker (right).





Exercises

- Additional (a bit more advanced) exercises on likelihood over pedigree data.

What's Next

- Chapter 4 - Aggregation, Heritability and Segregation Analysis