

# Module 6 - Aggregation, Heritability and Segregation Analysis

(Fundamentals of) Statistical Genetics

Lei Sun

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

## Chapter 4 - Aggregation, Heritability and Segregation Analysis

- ➡ Overview of aggregation and segregation (motivation)
- ➡ Aggregation analysis for binary trait
- ➡ Principle: shared genetic material among relatives
- ➡ IBS, IBD and Kinship Coefficient
- ➡ Recurrent risk ratio, link with disease model and estimation
- ➡ Example, cautions and pitfalls
- ➡ Heritability analysis for quantitative trait
- ➡ Additive model and General model
- ➡ Estimation using MZ and DZ twins
- ➡ Alternative approach to family aggregation analysis: case-control design.

# Outline II

- ➡ Segregation analysis
- ➡ Segregation ratio
- ➡ Autosomal dominant disease
- ➡ Autosomal recessive disease
- ➡ Beyond the simple models
- ➡ Design and sampling issues.

# Overview of Aggregation and Segregation Analyses I

**Before** conducting a genetic mapping studies of a trait of interest, we need to ask a couple of important questions.

- ➡ Is the trait worth genetic studies? Does the trait have **genetic basis**?
  - ◆ **Aggregation analyses for binary trait**
  - ◆ **Heritability analyses for quantitative traits**
  - ◆ *Designed to show that diseases, or phenotypes more generally, have a genetic basis by investigating patterns of **phenotypic correlation between relatives** (or clustering in families).*
  
- ➡ If so, what is the basic underlying **genetic model**?
  - ◆ **Segregation analysis** for simple Mendelian diseases.
  - ◆ *Used to find support for a specific genetic model underlying the inheritance patterns observed in families.*

# Overview of Aggregation and Segregation Analyses II

- *They all involve modelling phenotypic data on families, or pedigrees, **without using any genetic data**; all were developed during the time when genotyping was expensive, labor intensive, and not widely available.*
- *Coverage of these methods is brief: newer approaches are more popular, e.g. use population GWAS data (without pedigrees) to estimate heritability; use non-parametric linkage or just directly association analysis without the classical parametric linkage analysis.*
- **BUT**, *the concepts are useful to anyone with an interest in statistical genetics. In particular, 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.*

# Review of $P(Y, G)$ for Nuclear Families I

## 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,  $P_i \in \{0, 1, 2\}$ .
  
- ◆  $f(\cdot)$ : probability density function, e.g.  $P(X_1 = x_1) = f(x_1)$ .

## Joint Distribution of $Y$ and $G$ (Textbook Equation (4.3)).

$$\begin{aligned} f(y_1, y_2, x_1, x_2, g_1, g_2) &= f(y_1, y_2 | x_1, x_2, g_1, g_2) f(x_1, x_2, g_1, g_2) \\ &= f(y_1 | x_1) f(y_2 | x_2) f(x_1 | g_1, g_2) f(x_2 | g_1, g_2) f(g_1) f(g_2). \end{aligned}$$

# Review of $P(Y, G)$ for Nuclear Families II

- There are two major components in the probability formulation for  $G$ s of both parents and offsprings:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, P_1 = g_1, P_2 = g_2) \\ &= \frac{P(X_1 = x_1, X_2 = x_2 | P_1 = g_1, P_2 = g_2) P(P_1 = g_1, P_2 = g_2)}{1} \end{aligned}$$

- Parents generation,  $P(P_1, P_2)$ , needs concepts and models from population genetics (e.g. random mating, population allele frequency, HWE).

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

- 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).

$$\begin{aligned} f(x_1, x_2 | g_1, g_2) &= P(X_1 = x_1, X_2 = x_2 | P_1 = g_1, P_2 = g_2) \\ &= P(X_1 = x_1 | P_1 = g_1, P_2 = g_2) P(X_2 = x_2 | P_1 = g_1, P_2 = g_2) \\ &= f(x_1 | g_1, g_2) f(x_2 | g_1, g_2). \end{aligned}$$

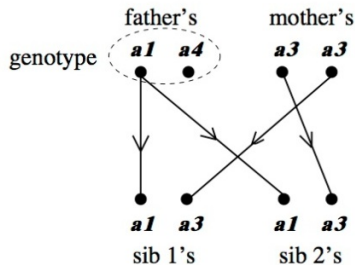
- The  $P(Y|X)$  component depends on the genetic models which typically specify the penetrance function.

# Principle of Aggregation Analysis

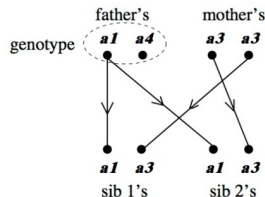
- ➡ If the phenotype of interest has a genetic component, the relative of an affected subject will have a higher predisposition to disease than an unrelated subject in the general population, because of the **shared genetic material among relatives**.

$$P(Y_{\text{Relative of } Y=1} | Y=1) > P(Y=1)$$

- ➡ Revisit Mendel's first law of segregation







➡ IBS: a set of alleles are said to be **Identical By State** if they are the same allelic type, e.g.

- ◆ *a1* of sib 1 and *a1* of sib 2
- ◆ *a3* of sib 1 and *a3* of sib 2

➡ IBD: a set of alleles are said to be **Identical By Descent** if they were inherited from the same ancestral allele (same origin), e.g.

- ◆ *a1* of sib 1 and *a1* of sib 2.

# IBD and Kinship Coefficient I

➡ IBD distribution,  $\vec{p} = (p_0, p_1, p_2)$ , for a pair of individuals

- ◆  $p_i = P(i \text{ number of alleles are shared IBD by a pair of individuals})$ .
- ◆ Can be used to **summarize and measure pairwise relationships**, e.g.

a MZ-twin pair:	$\vec{p} = (0, 0, 1)$
a full sib pair:	$\vec{p} = (1/4, 1/2, 1/4)$
a half-sib pair:	$\vec{p} = (1/2, 1/2, 0)$
a first-cousin pair:	$\vec{p} = (3/4, 1/4, 0)$
a unrelated pair:	$\vec{p} = (1, 0, 0)$

➡ Note that these IBD distributions are NOT conditional on any observed genotype data. They simply measure the probability that genetic material at a randomly selected locus from the genome to have common ancestry origin between two individuals. It is a characteristic for a population sample of specific relationship type, e.g. all full-sib pairs (not for one specific pair of interest.)

# IBD and Kinship Coefficient II

- ➡ Calculating the IBD distribution, e.g. half-sib pair,  $\vec{p} = (1/2, 1/2, 0)$ 
  - ◆ Let's assume the half-sib pair has the same father
  - ◆ Father's genotype is (a1, a4) as in the previous graph.
  - ◆ It's 'impossible' to share 2 alleles IBD, since there is no common origin from the two mother's side (unless the two mothers are relatives as well), so

$$p_2 = 0.$$

$$p_1 = P(1 \text{ IBD}) = P(\text{both inherited } a1) + P(\text{both inherited } a4) = \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{2} = \frac{1}{2}.$$

$$\begin{aligned} p_0 &= P(0 \text{ IBD}) = P(\text{sib1 inherited } a1 \text{ and sib2 inherited } a4) \\ &+ P(\text{sib1 inherited } a4 \text{ and sib2 inherited } a1) = \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{2} = \frac{1}{2}. \\ &(\text{ or } p_0 = 1 - p_1 - p_2 = 1 - \frac{1}{2} - 0 = \frac{1}{2}) \end{aligned}$$

- ◆ Note that the calculation does not depend on the specific genotype or how we code the genotype as long as we keep track the two origins/alleles. The two alleles could be the same allelic type (e.g. a3 and a3 of the mother in the previous graph, but they are two different origins in terms of IBD calculations.)

# IBD and Kinship Coefficient III

➡ Calculating the IBD distribution, e.g. full-sib pair,  $\vec{p} = (1/4, 1/2, 1/4)$ .

- ◆ Use the fact that full-sib pair is 'independent' sum of two half-sib pairs in terms of IBD sharing (Mendel's segregation from father to offsprings is independent of that from mother to offsprings.)

$$\begin{aligned}p_2 &= P(2 \text{ IBD}) = P(1 \text{ IBD from father side and 1 IBD from mother side}) \\&= P(1 \text{ IBD from father side})P(1 \text{ IBD from mother side}) = \frac{1}{2} \frac{1}{2} = \frac{1}{4}.\end{aligned}$$

$$\begin{aligned}p_0 &= P(0 \text{ IBD}) = P(0 \text{ IBD from father side and 0 IBD from mother side}) \\&= P(0 \text{ IBD from father side})P(0 \text{ IBD from mother side}) = \frac{1}{2} \frac{1}{2} = \frac{1}{4}.\end{aligned}$$

$$\begin{aligned}p_1 &= P(1 \text{ IBD}) = P(1 \text{ IBD from father side and 0 IBD from mother side}) \\&\quad + P(0 \text{ IBD from father side and 1 IBD from mother side}) \\&= P(1 \text{ IBD from father side})P(0 \text{ IBD from mother side}) \\&= P(0 \text{ IBD from father side})P(1 \text{ IBD from mother side}) \\&= \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

$$(\text{ or } p_1 = 1 - p_0 - p_2 = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2})$$

# IBD and Kinship Coefficient IV

- We emphasize here again that calculations of the previous IBD distributions do NOT conditional on any observed genotype data. It is a characteristic for a population sample of specific relationship type, e.g. all full-sib pairs.
- However, there are situations we might be interested in the IBD information for a particular pair.
- Assume genotypes of a pair of individuals are  $g_1 = (g_{1,1}, g_{1,2})$ ,  $g_2 = (g_{2,1}, g_{2,2})$ , e.g.  $g_1 = (a1, a3)$  and  $g_2 = (a1, a3)$ .
- Note that

$$P(0 \text{ IBD} \mid (a1, a3)(a1, a3) \text{ for a sib pair}) \neq 1/4!$$

- And this probability will further change if we add genotypes of the parents!
- And if we further assume we know the transition arrows? (Though this information is almost never known in practice; missing inheritance information.)
- We will come back to this later in non-parametric linkage analysis.












# IBD and Kinship Coefficient V

▀ Kinship Coefficient:  $\Phi$ , for a pair of individuals.

- ◆ Given two individuals  $i$  and  $j$ ,  $\Phi_{ij}$  is the probability that a randomly selected allele from individual  $i$ , and a randomly selected allele from individual  $j$  are IBD.
- ◆  $\Phi$  is actually a function of  $\vec{p} = (p_0, p_1, p_2)$ :

$$\begin{aligned}\Phi &= P(\text{the two randomly selected alleles are IBD}) \\ &= P(\text{the two randomly selected alleles are IBD} \mid 1 \text{ allele IBD})p_1 \\ &\quad + P(\text{the two randomly selected alleles are IBD} \mid 2 \text{ alleles IBD})p_2 \\ &= \frac{1}{4}p_1 + \frac{1}{2}p_2.\end{aligned}$$

- ◆  $\Phi$  can also be used to summarize and measure pairwise relationships (but with only 1 d.f.).

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

Relationship type (notation)	Distribution of IBD Sharing			Kinship coefficient, $\phi$
	$p_0$	$p_1$	$p_2$	
MZ-twin (MZ)	0	0	1	0.5
Parent-offspring (PO)	0	1	0	0.25
Full-sib (FS)	0.25	0.5	0.25	0.25
Half-sib + first-cousin (HSFC)	0.375	0.5	0.125	0.1875
Half-sib (HS)	0.5	0.5	0	0.125
Grandparent-grandchild (GPC)	0.5	0.5	0	0.125
Avuncular (AV)	0.5	0.5	0	0.125
First-cousin (FC)	0.75	0.25	0	0.0625
Half-avuncular (HAV)	0.75	0.25	0	0.0625
Half-first-cousin (HFC)	0.875	0.125	0	0.03125
Unrelated (UN)	1	0	0	0

[Detecting pedigree relationship errors](#) In Statistical Human Genetics: Methods and Protocols. Elston R, Satagopan J and Sun S (editors), Human Press, Inc. Springer, pp.25 to 46.



# Principle of Aggregation Analysis - Example I

- ➡ *If the phenotype of interest has a genetic component, the relative of an affected subject will have a higher predisposition to disease than an unrelated subject in the general population, because of the **shared genetic material among relatives**.*

$$P(Y_{\text{Relative of } Y=1} = 1 | Y = 1) > P(Y = 1)$$

- ➡ Let's assume

- ◆ A very simple Mendelian recessive model of  $f_0 = f_1 = 0$  and  $f_2 = 1$ .
- ◆ Note that  $Y = 1$  implies genotype is  $DD$ .
- ◆ The relationship between two individuals is full-sib.
- ◆ Allele frequency of  $D$  is  $p$ .

# Principle of Aggregation Analysis - Example II

► We can show that  $P(Y_{\text{Sib of } Y} = 1 | Y = 1) > P(Y = 1)$ .

$$P(Y = 1) = P(DD) = p^2.$$

$$P(Y_2 = 1 | Y_1 = 1) = \frac{P(Y_1 = 1, Y_2 = 1)}{P(Y_1 = 1)}$$

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1) &= P(X_1 = DD, X_2 = DD) \\ &= P(X_1 = DD, X_2 = DD, P_1 = DD, P_2 = DD) \\ &\quad + P(X_1 = DD, X_2 = DD, P_1 = DD, P_2 = Dd) \\ &\quad + P(X_1 = DD, X_2 = DD, P_1 = Dd, P_2 = DD) \\ &\quad + P(X_1 = DD, X_2 = DD, P_1 = Dd, P_2 = Dd) \\ &= 1 \cdot 1 \cdot p^2 \cdot p^2 + 2\left(\frac{1}{2} \cdot \frac{1}{2} \cdot p^2 \cdot 2p(1-p)\right) + \frac{1}{4} \cdot \frac{1}{4} \cdot 2p(1-p) \cdot 2p(1-p) \\ &= p^4 + p^3(1-p) + \frac{1}{4}p^2(1-p)^2. \end{aligned}$$

# Principle of Aggregation Analysis - Example III

$$\begin{aligned} P(Y_2 = 1 | Y_1 = 1) &= \frac{P(Y_1 = 1, Y_2 = 1)}{P(Y_1 = 1)} \\ &= \frac{p^4 + p^3(1-p) + \frac{1}{4}p^2(1-p)^2}{p^2} = p^2 + p(1-p) + \frac{1}{4}(1-p)^2 \\ &= p^2(1 + \frac{(1+3p)(1-p)}{4p^2}) > p^2 = P(Y_1 = 1). \end{aligned}$$

- Although we did not see the IBD sharing entering into the calculation directly, in this simple case, we can intuitively understand that  $P(X_2 = DD | X_1 = DD) > P(X = DD)$  if the two individuals are relatives, because genotype of one individual informs the other due to the common ancestry (potential IBD sharing).
- Note that the recessive disease with full penetrance assumption makes the calculation much easier. If we assume a dominant disease, then there would be more summations to consider. We can also perform the calculation for other relative pairs, e.g. parent-offspring?
- Now let's formally define a quantity (recurrent risk ratio) that measures the strength of the genetic aggregation among relatives.

# Recurrent Risk Ratio for Binary Traits

- Y: a binary trait of interest,  $Y = 1$  denotes a person being affected.
- K: population prevalence  $K = P(Y = 1)$ . (Note the simplification of the model: absence of other covariates such as age and sex etc.)
- $Y_1$  and  $Y_2$ : phenotypes of two individuals.
- R: the relationship between two individuals, e.g. MZ twins, DZ twins, siblings, half-sibs etc.
- Recurrent risk ratio** measures the strength of the genetic aggregation among relatives.

$$\lambda_R = \frac{P(Y_2 = 1 | Y_1 = 1)}{K}.$$

- If the disease has genetic basis, we would expect  $\lambda_R \nearrow$  as the relationship R gets 'closer', e.g. Textbook Table 4.1.

**Table 4.1** Observed recurrence risk ratios from a sample of families with schizophrenia. *Source: Risch (1990a)*

Risk Ratio	$\lambda_O$	$\lambda_S$	$\lambda_M$	$\lambda_D$	$\lambda_H$	$\lambda_N$	$\lambda_G$	$\lambda_C$
Observed	10.0	8.6	52.1	14.2	3.5	3.1	3.3	1.8

Definitions of subscripts: O = offspring; S = sibling; M = MZ twins; D = DZ twins; H = half-sibs; N = niece/nephew; G = grandchild; C = first cousins.

# Recurrent Risk Ratio - Cautions

- Question: does  $\lambda_R > 1$  prove that the disease has genetic basis?
- Cautions (relevant for both binary and continuous traits): **due to shared exposure to similar environment, it's possible that a disease having no genetic etiology could also show evidence of familial clustering.**
- e.g. flue, infectious disease:  $P(Y_2 = 1|Y_1 = 1)$  for relatives is most likely greater than the population prevalence  $K$ , so  $\lambda_R > 1$ .  
In fact,  $\lambda_S$  is most likely greater than  $\lambda_C$  as well because of the greater amount of shared environment.

# Recurrent Risk Ratio and Disease Models I

- ➡ Linking the quantities that we have learned so far.
- ➡ Get more familiar with the Bayes' rule and conditional probabilities.
- ➡ For the simple sibling case and assume simple Mendelian disease.

$$\lambda_R = \lambda_S = \frac{P(Y_2 = 1|Y_1 = 1)}{K} = \frac{P(Y_2 = 1, Y_1 = 1)}{P(Y_1 = 1)K} = \frac{P(Y_2 = 1, Y_1 = 1)}{P(Y = 1)^2}.$$

- ➡ Denominator (Textbook Equation 4.6)

$$\begin{aligned} K = P(Y = 1) &= \sum_{x=0,1,2, \text{ or } dd, dD, DD} P(Y = 1, X = x) \\ &= \sum_{x=0,1,2} P(Y = 1|X = x)P(X = x) = f_0(1-p)^2 + f_1 2p(1-p) + f_2 p^2. \end{aligned}$$

(In our previous example,  $f_0 = f_1 = 0$  and  $f_2 = 1$ .)

# Recurrent Risk Ratio and Disease Models II

## ➡ Numerator (Textbook Equation 4.5)

$$\begin{aligned} & P(Y_2 = 1, Y_1 = 1) \\ &= \sum_{x_1, x_2, g_1, g_2 \in \{0,1,2\}} P(Y_2 = 1, Y_1 = 1, X_1 = x_1, X_2 = x_2, P_1 = g_1, P_2 = g_2) \\ & \quad \text{(use Equation 4.3)} \\ &= \sum_{x_1, x_2, g_1, g_2 \in \{0,1,2\}} f(y_1|x_1)f(y_2|x_2)f(x_1|g_1, g_2)f(x_2|g_1, g_2)f(g_1)f(g_2) \\ &= \sum_{x_1, x_2, g_1, g_2 \in \{0,1,2\}} f_{x_1} f_{x_2} f(x_1|g_1, g_2)f(x_2|g_1, g_2)f(g_1)f(g_2) \\ &= \sum_{g_1, g_2 \in \{0,1,2\}} f(g_1)f(g_2) \left\{ \sum_{x_1 \in \{0,1,2\}} f_{x_1} f(x_1|g_1, g_2) \sum_{x_2 \in \{0,1,2\}} f_{x_2} f(x_2|g_1, g_2) \right\}. \end{aligned}$$

(In our previous example, the affected two sibs must have genotype *DD*, so  $x_1 = x_2 = 2$ . In return, both parents must carry at least one copy of *D*, so  $g_1 \neq 0$  and  $g_2 \neq 0$ . All these make the number of summations substantially smaller.)

# Recurrent Risk Ratio - Estimation I

- For the simple case of siblings:  $\lambda_S$ .
- Obtain a sample of unrelated (and matched) cases and controls (called probands).
- Obtain clinical diagnoses of the siblings of the cases and controls.
- Calculate  $s_{case}$ , the proportion of affected siblings among siblings of the cases.

$$s_{case} \approx P(Y_2 | Y_1).$$

$$\hat{\lambda}_S = \frac{s_{case}}{K}.$$

- If the absence of  $K$ , calculate  $s_{controls}$ , the proportion of affected siblings among siblings of the controls. If the disease is rare,

$$s_{controls} \approx K = P(Y = 1).$$



# Recurrent Risk Ratio - Estimation II

- ➡ Many pitfalls and complications: matching cases and controls, adjusting for ascertainment bias, and accounting for covariates such as age at onset or environmental factors that required more advanced stat gene training, e.g.
  - ◆ Guo (1998). [Inflation of Sibling Recurrence-Risk Ratio, Due to Ascertainment Bias and/or Overreporting](#). American Journal of Human Genetics.
  - ◆ Liang and Beaty (2010). [Statistical designs for familial aggregation](#). Statistical Methods in Medical Research.
  - ◆ Javaras et al. (2010). [Estimating Disease Prevalence Using Relatives of Case and Control Probands](#). Biometrics.
- ➡ We will skip discussion on Attributable Fraction which assesses the genetic effect relative to the disease prevalence.

$$AF = 1 - \frac{P(Y = 1|dd)}{P(Y = 1)} = 1 - \frac{f_0}{K}.$$

# Alternative Approaches to Familial Aggregation I

## Use the conventional case-control design to detect familial aggregation:

- Idea: compare the prevalence of family history of the trait/disease of interest (e.g. breast cancer) between cases (affected) and controls (unaffected)

$$P(\text{positive family history} | \text{cases}) > P(\text{positive family history} | \text{controls})?$$

- Data: of a random sample of cases and controls and their family history can be summarized as the classical 2x2 table, e.g.
- a genetic study of chronic obstructive pulmonary disease by Cohen (1980): 105 cases and 79 controls were sampled from the Johns Hopkins Hospital:

	Family History		Total
	Positive	Negative	
Cases	50 ( $n_{11}$ )	55 ( $n_{12}$ )	105 ( $n_{1.}$ )
Controls	23 ( $n_{21}$ )	56 ( $n_{22}$ )	79 ( $n_{2.}$ )
Total	73 ( $n_{.1}$ )	111 ( $n_{.2}$ )	184 ( $n_{..}$ )

- Hypothesis/question of interest: is there any association/relationship between affection status and family history?

# Alternative Approaches to Familial Aggregation II

## Also a number of design issues/pitfalls

### ➡ What is positive family history?

It is usually defined as the presence of disease in one or more first-degree relatives for either the cases or controls. (e.g. if the mother or sisters have breast cancer.) Would depend on family size, degree of relatedness, age of the relatives etc.

### ➡ How do you obtain the information?

Usually obtained by interviewing cases and controls themselves, or through family members. Again, there might be potential bias of information of recall (the true disease status of relatives may be misreported), depending on the biological relationship, number of relatives interviewed and other factors. (Use of parents or spouses as informants, and multiple informants may be desirable.

### ➡ Case-control matching cannot be overstated: some important confounding or external factors should be comparable between the sample of cases and controls, including family size, biological relationships of relatives to the subject, and the age distribution of these relatives.

For example, a subject with 5 adult sisters is more likely to have positive family history of breast cancer than a subject with 1 teen-age sister.

# Alternative Approaches to Familial Aggregation III

## Analysis issues

- Are we interested in comparing

$P(\text{positive family history} \mid \text{cases})$  with  $P(\text{positive family history} \mid \text{controls})$

or

$P(\text{cases} \mid \text{positive family history})$  with  $P(\text{cases} \mid \text{negative family history})$ ?

- But can we still use the case-control study design for the latter?

- The use of **Odds Ratio**

$$\frac{\text{odds}(\text{cases} \mid \text{positive FH})}{\text{odds}(\text{cases} \mid \text{negative FH})} = \frac{\text{odds}(\text{positive FH} \mid \text{cases})}{\text{odds}(\text{positive FH} \mid \text{controls})}$$

- We defer OR discussion to population-based association analysis.

# Heritability - Additive Model I

- ➡ *We assume here that the trait of interest is measured on a quantitative scale, i.e., height, weight, blood pressure, etc.*
- ➡ *Heritability analysis assesses the overall genetic contribution to the variation in the phenotype.*
- ➡ Let's consider the simple additive normal model,

$$Y = \alpha + \beta G + e,$$

where  $e$  captures the effect of environmental (non-genetic) factors.

- ➡ What is the total variation in the phenotype  $Y$ ?

$$\text{Var}(Y) = \beta^2 \text{Var}(G) + \text{Var}(e) + 2\text{Cov}(G, e).$$

# Heritability - Additive Model II

$$\text{Var}(Y) = \beta^2 \text{Var}(G) + \text{Var}(e) + 2\text{Cov}(G, e).$$

- ▀ If we assume  $G$  and  $e$  are independent (*not true in general, but it is a reasonable hypothesis in heritability analysis where  $\text{Var}(G) \gg \text{Cov}(G, e)$* ).

- ▀ Then we have the following

$$\text{Var}(Y) = \beta^2 \text{Var}(G) + \text{Var}(e)$$

- ▀ Complex traits are typically affected by multiple genes, so

$$Y = \alpha + \sum_m \beta_m G_m + e,$$

$$\text{Var}(Y) = \sum_m \beta_m^2 \text{Var}(G_m) + \text{Var}(e).$$

# Heritability - Additive Model III

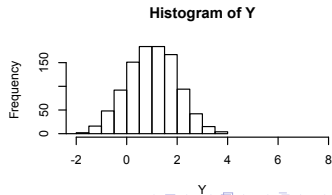
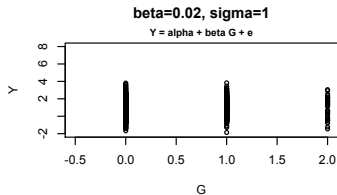
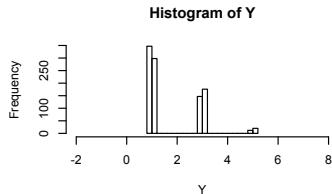
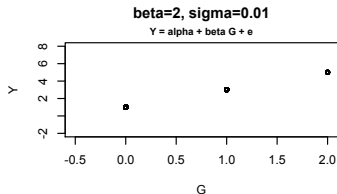
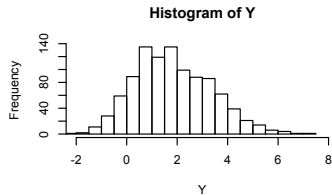
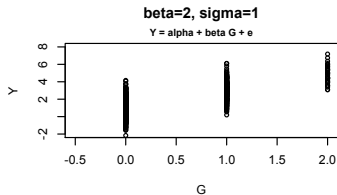
- Often we use the following notation to denote the **variance partition**:

$$V_Y = V_G + V_E.$$

- A natural choice of measure of the genetic contribution (heritability) would be

$$h^2 = \frac{V_G}{V_Y} = \frac{V_G}{V_G + V_E}.$$

- Graphic illustration of the variance partition. [R codes](#)





# Heritability - A General Model Framework

- Now let's consider a more general model using a slightly different notation as before to be consistent with the Textbook. We code  $G = 0, 1$  and  $2$  'additively' for  $dd, Dd$  and  $DD$ , but use the following (2 d.f.) model.

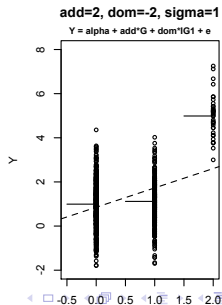
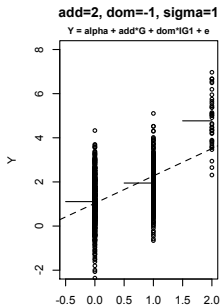
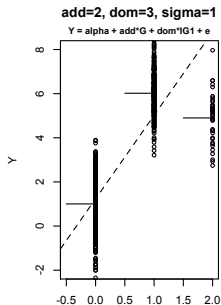
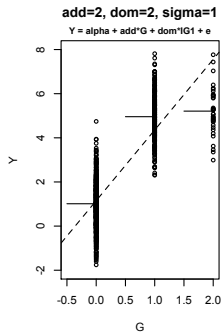
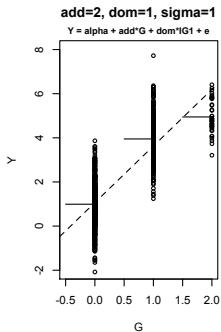
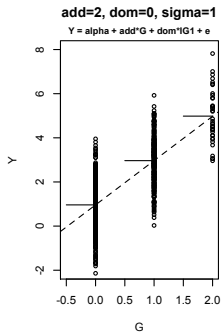
$$Y = \mu + aG + dI(G = 1) + e.$$

$$(Y|G = dd) \sim N(\mu, \sigma^2), \quad \mu_0 = \mu,$$

$$(Y|G = dD) \sim N(\mu + a + d, \sigma^2), \quad \mu_1 = \mu + a + d,$$

$$(Y|G = DD) \sim N(\mu + 2a, \sigma^2), \quad \mu_2 = \mu + 2a.$$

- Now putting different constraints on  $d$  would lead to different models, e.g.
- ◆ Recessive:  $d = -a$
  - ◆ Dominant:  $d = a$
  - ◆ Additive:  $d = 0$



# Heritability - General Model I

- ➡ Note that with  $Y = \mu + aG + dI(G = 1) + e$ , we have

$$\text{Var}(Y) = a^2 \text{Var}(G) + d^2 \text{Var}(I(G = 1)) + 2ad \text{Cov}(G, I(G = 1)) + \sigma^2.$$

- ➡ We can still use the following notation to denote the variance partition:

$$V_Y = V_G + V_E$$

- ➡ But, *the genetic variance can be [further] partitioned into the Additive Genetic Variance  $V_A$  and the Dominant Genetic Variance  $V_G$  (Falconer and Mackay (1996)).*

$$V_G = V_A + V_D.$$

- ➡ Then we will have two definitions of heritability:

the broad sense heritability  $h^2 = \frac{V_G}{V_Y}$

the narrow sense heritability  $h^2 = \frac{V_A}{V_Y}$

# Heritability - General Model II

- Some detailed derivation.

$G =$	$I(G) =$	with Probability
0	0	$(1 - p)^2$
1	1	$2p(1 - p)$
2	0	$p^2$

- Mean and variance of  $G$ .

$$E(G) = 2p(1 - p) + 2p^2 = 2p.$$

$$\text{Var}(G) = E(G^2) - (E(G))^2 = 2p(1 - p) + 4p^2 - (2p)^2 = 2p(1 - p).$$

- Mean and variance of the indicator variable  $I = I(G = 1)$ .

$$E(I) = 2p(1 - p).$$

$$\text{Var}(I) = E(I^2) - (E(I))^2 = 2p(1 - p) - (2p(1 - p))^2.$$

- Covariance between  $G$  and  $I$ .

$$\text{Cov}(G, I) = E(GI) - E(G)E(I) = 2p(1 - p) - 2p \cdot 2p(1 - p) = 2p(1 - p)(1 - 2p).$$

# Heritability - General Model III

$$\text{Var}(Y) = a^2 \text{Var}(G) + d^2 \text{Var}(I(G = 1)) + 2ad\text{Cov}(G, I(G = 1)) + \sigma^2.$$

$$\begin{aligned} V_G &= a^2 \text{Var}(G) + d^2 \text{Var}(I(G = 1)) + 2ad\text{Cov}(G, I(G = 1)) \\ &= a^2 2p(1-p) + d^2 (2p(1-p) - (2p(1-p))^2) + 2ad 2p(1-p)(1-2p) \\ &= 2p(1-p) (a^2 + 2ad(1-2p) + d^2(1-2p(1-p))) \\ &= 2p(1-p) (a^2 + 2ad(1-2p) + d^2(1-2p)^2) \\ &\quad + 2p(1-p) (d^2(1-2p(1-p)) - d^2(1-2p)^2) \\ &= \frac{2p(1-p)(a + d(1-2p))^2}{1} + \frac{(2p(1-p)d)^2}{1} \\ &= V_A + V_D. \end{aligned}$$

➡ This is the Equation 4.8 of the Textbook, without the subscript  $m$  to indicate multiple Gs.

➡ For the simple additive model where  $d = 0$ , then

$$V_G = V_A = 2p(1-p)a^2$$

# Heritability Estimation

- ➡ If we were able to obtain homozygous lines from the population (possible in animal and plant genetics),  
then  $V_E$  can be estimated from the homozygous subpopulation and  $V_Y$  estimated from the over all population, then  $\hat{h}^2 = (\hat{V}_Y - \hat{V}_E) / \hat{V}_Y$ .
- ➡ Heritability can be directly estimated from the phenotypic data on **relatives**, by cleverly linking the **phenotypic correlation** with the  $V_A$  component.  
e.g. see Box 4.1 of the Textbook on Estimation of  $h^2$  using Parent-Child trios.
- ➡ Caution: estimates based on correlations between relatives *depend critically on the assumption that environmental correlations between individuals are the same for all degrees of relationship.*  
*If closer relatives have more similar environments, as they do in humans, the estimates of heritability are biased.*

# Heritability Estimation - MZ vs. DZ Twin Design I

The difference in phenotypic correlation between monozygotic (MZ) and dizygotic (DZ) twins is also commonly used in human genetics to estimate  $h^2$ .

▀ Let's consider the normal additive model

$$Y_1 = \mu + aG_1 + e_1, \quad Y_2 = \mu + aG_2 + e_2.$$

▀ For MZ twins,  $G_1 = G_2$ , so

$$\text{Cov}(G_1, G_2) = \text{Var}(G) = 2p(1 - p).$$

▀ The phenotypic covariance is then

$$\text{Cov}(Y_1, Y_2) = a^2 \text{Cov}(G_1, G_2) = a^2 2p(1 - p) = V_A.$$

# Heritability Estimation - MZ vs. DZ Twin Design II

- Because  $Var(Y_1) = Var(Y_2)$ , so

$$Var(Y) = \sqrt{Var(Y_1)}\sqrt{Var(Y_2)} = V_Y.$$

- Thus

$$h^2 = \frac{V_A}{V_Y} = \frac{Cov(Y_1, Y_2)}{\sqrt{Var(Y_1)}\sqrt{Var(Y_2)}} = Corr(Y_1, Y_2) = \rho_{MZ}.$$

- So, why not just collect a sample of MZ twins and using sample estimate of the correlation between  $Y_1$  and  $Y_2$  as the estimate for  $h^2$ ?



# Heritability Estimation - MZ vs. DZ Twin Design III

$$Y_1 = \mu + aG_1 + e_1, \quad Y_2 = \mu + aG_2 + e_2.$$

- ➡  $Cov(Y_1, Y_2) = a^2 Cov(G_1, G_2)$  assumed that  $e_1$  and  $e_2$  are independent of each other (as well as independent of  $G_1$  and  $G_2$ ). However,  $e_1$  and  $e_2$  independence is highly unlikely, so in fact,

$$Cov(Y_1, Y_2|MZ) = a^2 Cov(G_1, G_2|MZ) + Cov(e_1, e_2|MZ) = \underline{V_A + Cov(e_1, e_2|MZ)}.$$

- ➡ What is  $Cov(G_1, G_2)$  for DZ twins (genetically they are siblings)? It turns out

$$Cov(G_1, G_2|DZ) = p(1 - p).$$

- ➡ Thus, for DZ twins we have

$$Cov(Y_1, Y_2|DZ) = a^2 Cov(G_1, G_2|DZ) + Cov(e_1, e_2|DZ).$$

$$= a^2 p(1 - p) + Cov(e_1, e_2|DZ) = \underline{\frac{V_A}{2} + Cov(e_1, e_2|DZ)}.$$

# Heritability Estimation - MZ vs. DZ Twin Design IV

- It is reasonable to assume  $Cov(e_1, e_2|MZ) \approx Cov(e_1, e_2|DZ)$ , so

$$\begin{aligned}\rho_{MZ} - \rho_{DZ} &= Corr(Y_1, Y_2|MZ) - Corr(Y_1, Y_2|DZ) \\&= \frac{Cov(Y_1, Y_2|MZ)}{\sqrt{Var(Y_1)}\sqrt{Var(Y_2)}} - \frac{Cov(Y_1, Y_2|DZ)}{\sqrt{Var(Y_1)}\sqrt{Var(Y_2)}} \\&= \frac{V_A + Cov(e_1, e_2|MZ)}{V_Y} - \frac{\frac{V_A}{2} + Cov(e_1, e_2|DZ)}{V_Y} \\&= \frac{1}{2} \frac{V_A}{V_Y} = \frac{1}{2} h^2\end{aligned}$$

- Thus we can contrast the sample estimates of phenotypic correlation between MZ and DZ twins to estimate the  $h^2$ ,  $\hat{h}^2 = 2(\hat{\rho}_{MZ} - \hat{\rho}_{DZ})$ .

# Heritability Estimation - MZ vs. DZ Twin Design V

- ➡ Note that if we replaced DZ twins with siblings then the assumption  $Cov(e_1, e_2|MZ) \approx Cov(e_1, e_2|sibs)$  would be less reasonable; hence we used MZ and DZ design even if  $Cov(G_1, G_2|DZ) = Cov(G_1, G_2|sibs)$ !
- ➡ Also note that textbook example of estimating  $h^2$  using Parent-Child trios assumes that  $E$  of child and average  $E$  of parents are uncorrelated.
- ➡ It also involves calculation of the correlation between genotype of offspring  $G_o$  and average genotype of the parents  $G_p$ , which turned out to be  $Cov(G_o, G_p) = p(1 - p)$ .

# Details of the $Cov(G_1, G_2)$ Calculation I

- How did we get  $Cov(G_1, G_2) = p(1 - p)$  for a DZ twin pair?  
(Note that DZ=sib pair genetically.)

(Unordered) genotype	$G_1 \cdot G_2 =$	with Probability
dd dd	0	NA
dd dD	0	NA
dd DD	0	NA
dD dD	1	$p^2(1 - p)^2 + p(1 - p)$
dD DD	2	$p^3(1 - p) + p^2(1 - p)$
DD DD	4	$\frac{1}{4}p^4 + \frac{1}{2}p^3 + \frac{1}{4}p^2$

$$Cov(G_1, G_2) = E(G_1 \cdot G_2) - E(G_1)E(G_2) = E(G_1 \cdot G_2) - (2p)^2.$$

$$\begin{aligned} E(G_1 \cdot G_2) &= p^2(1 - p)^2 + p(1 - p) + 2(p^3(1 - p) + p^2(1 - p)) + 4(\frac{1}{4}p^4 + \frac{1}{2}p^3 + \frac{1}{4}p^2) \\ &= 3p^2 + p. \end{aligned}$$

$$Cov(G_1, G_2) = 3p^2 + p - (2p)^2 = p(1 - p).$$

# Details of the $\text{Cov}(G_1, G_2)$ Calculation II

- ➡ To calculate  $P(G_1, G_2)$  for the siblings, we can use the technique of summation over all six possible (unordered) parental genotypes that we have learned.
- ➡ Alternatively, we can use a less tedious technique by summation over the three IBD status.
- ➡ To do this, we need the following conditional probability  $P(G_1, G_2 | \text{IBD})$ . Note that this conditional probability does not depend on the relationship type.

Unordered genotype	IBD status		
	0	1	2
$(ii, ii)$	$f_i^4$	$f_i^3$	$f_i^2$
$(ii, ij)$	$4f_i^3 f_j$	$2f_i^2 f_j$	0
$(ii, jj)$	$2f_i^2 f_j^2$	0	0
$(ii, jk)$	$4f_i^2 f_j f_k$	0	0
$(ij, ij)$	$4f_i^2 f_j^2$	$f_i f_j (f_i + f_j)$	$2f_i f_j$
$(ij, ik)$	$8f_i^2 f_j f_k$	$2f_i f_j f_k$	0
$(ij, kl)$	$8f_i f_j f_k f_l$	0	0

**Detecting pedigree relationship errors** In Statistical Human Genetics: Methods and Protocols. Elston R, Satagopan J and Sun S (editors), Human Press, Inc. Springer, pp.25 to 46.

# Details of the $\text{Cov}(G_1, G_2)$ Calculation III

► For example, for a pair of sib pairs or DZ twin pair,

$$\begin{aligned} & P(dD, dD) \\ = & P(dD, dD|0 \text{ IBD})P(0 \text{ IBD}) + P(dD, dD|1 \text{ IBD})P(1 \text{ IBD}) + P(dD, dD|2 \text{ IBD})P(2 \text{ IBD}) \\ = & 4p^2(1-p)^2\frac{1}{4} + p(1-p)\frac{1}{2} + 2p(1-p)\frac{1}{4} = p^2(1-p)^2 + p(1-p). \end{aligned}$$

$$P(dD, DD) = 4p^3(1-p)\frac{1}{4} + 2p^2(1-p)\frac{1}{2} = p^3(1-p) + p^2(1-p).$$

$$P(DD, DD) = p^4\frac{1}{4} + p^3\frac{1}{2} + p^2\frac{1}{4}.$$

# Heritability - Additional Considerations I

- ➡ Recently, new approaches have been proposed using **population 'unrelated' individuals**, by Dr. Peter Visscher's group.

*The fundamental idea is to estimate the heritability due to common variants by studying the extent to which the phenotypic similarity across pairs of individuals in a sample is explained by their genotypic similarity at common variants. Rather than using simple correlation, they used a family of elegant statistical models, called linear mixed models (LMMs), and estimated the heritability using a technique called restricted maximum likelihood (REML) estimation.*

- ➡ **But cautions and pitfalls remain!**
- ➡ *these methods [LMMs and REML] seriously underestimate the true heritability when applied to case-control studies of disease..*
- ➡ Genetic interaction can lead to overestimation of heritability.

# Heritability - Additional Considerations II

- ➡ The issue of **missing heritability** is a big one!
- ➡ **Personal genomes: The case of the missing heritability.** *When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. Brendan Maher shines a light on six places where the missing loot could be stashed away.*
- ➡ **Missing heritability and strategies for finding the underlying causes of complex disease.** *Although recent genome-wide studies have provided valuable insights into the genetic basis of human disease, they have explained relatively little of the heritability of most complex traits, and the variants identified through these studies have small effect sizes. This has led to the important and hotly debated issue of where the 'missing heritability' of complex diseases might be found. Here, seven leading geneticists offer their opinion about where this heritability is likely to lie, what this could tell us about the underlying genetic architecture of common diseases and how this could inform research strategies for uncovering genetic risk factors.*



# Segregation Analysis - Segregation Ratio

- ➡ **Segregation ratios:** The proportions of the different genotypes and phenotypes in the offspring of the 6 parental mating types.
- ➡ e.g. segregation ratios for an autosomal dominant disease:

6 parental mating type	Offspring Genotype (in probability)			Offspring Phenotype (in probability)	
	DD	Dd	dd	Affected	Normal
DD x DD	1	0	0	1	0
DD x Dd	1/2	1/2	0	1	0
DD x dd	0	1	0	1	0
Dd x Dd	1/4	1/2	1/4	3/4	1/4
Dd x dd	0	1/2	1/2	1/2	1/2
dd x dd	0	0	1	0	1

- ➡ The goal of segregation analysis is to determine whether segregation ratios are consistent with expectations of autosomal dominant or recessive transmission.
- ➡ The demonstration of such ratios gives strong evidence that the trait under the study has a simple genetic basis.

# Segregation Analysis - Autosomal Dominant Disease I

## Underlying model and assumptions

- ◆  $D$  is the mutant allele and  $d$  is the normal allele
- ◆ Autosomal dominant diseases are usually rare, which implies that the allele frequency of allele  $D$  is low, i.e.  $D$  is rare.
- ◆ Thus, a random sample of an affected individual is most likely to carry the  $dD$  genotype. (Remember that the frequency of genotype  $DD$  is  $p^2$ .)

$$p(DD|affected) = \frac{p^2}{p^2 + 2p(1-p)^2} = \frac{p}{2-p} \approx \frac{p}{2}.$$

- ◆ **Design:** use a random sample of matings between affected (assumed to have genotype  $dD$ ) and unaffected individuals ( $dd$ ).
- ◆ **Data:** observe  $n$  offsprings in total, among which  $n_{Affected}$  offsprings are affected by the disease.

# Segregation Analysis - Autosomal Dominant Disease II

## Questions of interest

- ◆ Estimation: what is the segregation ratio  $p$ ?
- ◆ For autosomal dominant disease, an offspring of mating type  $dD \times dd$  has probability  $p = 1/2$  of being affected.
- ◆ Hypothesis testing: can we reject the  $H_0$  that  $p = p_0 = 1/2$ ?

It's a simple problem, but again we will use this example to exercise different estimation and testing methods, and to further grasp some of the basic statistical principles.

# Segregation Analysis - Autosomal Dominant Disease III

- ➡ First note the **Binomial distribution**:

$$n_{Affected} \sim Bino(n, p)$$

- ➡ Important Question to ask: two affected sibs are independent of each other?
- ◆ Mendelian transition from parents to one sib is independent of that of the transition to another sib.
  - ◆ **However**, if there are contributing covariates, then affected siblings are not independent due to common shared environmental effect.

- ➡ **Estimation - MLE** (the same principle as the allele frequency estimation).

$$\hat{p} = \frac{n_{Affected}}{n}.$$

# Segregation Analysis - Autosomal Dominant Disease IV

## Method of Moment (CHL 5224)

- ◆ Express the theoretical values of the moments (e.g the first moment) of certain random variable in terms of the parameters to be estimated.
- ◆ Then equates these theoretical equations to the values of these moments calculated from the observed data, and solve the “estimating equations” to obtain the estimates.
- ◆ In this case
  - \* random variable:  $r$ , the number of affected out of  $n$ ,
  - \* parameter of interest:  $p$ , the segregation ratio.
- ◆ The first moment of  $r$  is

$$E[r] = n \cdot p$$

$$\implies p = \frac{r}{n}$$

# Segregation Analysis - Autosomal Dominant Disease V

- ➡ **Hypothesis testing - Likelihood Ratio Test** (the same principle as the allele frequency testing).

$$H_0 : p = p_0 = \frac{1}{2} \text{ (Dominant mode of inheritance).}$$

$$\begin{aligned} T &= 2(l(\hat{p}) - l(\tilde{p})) = 2(l(\hat{p}) - l(p_0)) = 2 \sum \text{observed} \times \log \frac{\text{observed}}{\text{expected}} \\ &= 2 \left( n_{\text{Affected}} \log\left(\frac{n_{\text{Affected}}}{np_0}\right) + (n - n_{\text{Affected}}) \log\left(\frac{n - n_{\text{Affected}}}{n(1 - p_0)}\right) \right) \\ &= 2 \left( n_{\text{Affected}} \log\left(\frac{\hat{p}}{1/2}\right) + (n - n_{\text{Affected}}) \log\left(\frac{1 - \hat{p}}{1/2}\right) \right) \sim \chi_1^2 \end{aligned}$$

This is the SAME (a slightly condensed version) as the expression shown on page 60 of the Textbook.

# Segregation Analysis - Autosomal Dominant Disease VI

## Other tests (CHL 5224)

- Binomial exact test.

$$\text{p-value} = 2P(r \geq r_{obs}) \text{ if } r_{obs} \geq n/2, \text{ or } = 2P(r \leq r_{obs}) \text{ if } r_{obs} < n/2.$$

- Normal approximation to Binomial test (with or without continuity correction).

$$r \sim N(np, np(1-p)).$$

- Pearson  $\chi_r^2$  test.

- A few notes on likelihood ratio test and Pearson  $\chi_r^2$  test.

- ◆ The proof of  $X = 2\ln\left\{\frac{L_{H_1}(\hat{\theta})}{L_{H_0}(\hat{\theta})}\right\} \approx \chi_r^2$  is based on the Taylor's expansion w.r.t.  $\theta$ .
- ◆ Pearson  $\chi^2$  test is a large sample approximation to  $2\ln\lambda$ , an approximation which depends only on the restricted MLE of  $\theta$  under the null hypothesis. This may be easier to calculate than LRT which requires unrestricted MLE. However, in many complex situations, only likelihood approach is applicable.
- ◆ Subject to regularity conditions, the two tests have approximately the same power function for large samples (large-sample equivalence). In that case, we may choose the test that is most convenient computationally.

# Segregation Analysis - Autosomal Dominant Disease VII

➡ **The Sickling trait example** (page 60 and Figure 4.3 of the Text book):

$n = 23$  offsprings from 4 matings and  $n_{Affected} = 11$ .

◆ MLE is

$$\hat{p} = \frac{n_{Affected}}{n} = \frac{11}{23} = 0.478.$$

◆ p-value of the likelihood ratio test is

$$P(T \geq T_{obs}) = P(T \geq 0.0435) = 0.83.$$

◆ No evidence to reject the null.



# Segregation Analysis - Autosomal Dominant Disease VIII

➡ What if  $n = 2300$  and  $n_{Affected} = 1100$ ?

- ◆ MLE is  $\hat{p} = \frac{n_{Affected}}{n} = \frac{1100}{2300} = 0.478$ , identical as above.
- ◆ p-value of the likelihood ratio test is  $P(T \geq T_{obs}) = P(T \geq 4.35) = 0.037$
- ◆ Now we have evidence to reject the null.
- ◆ Important reminder: think about the **variance** associated with  $\hat{p}$ .

➡ We can also perform the other tests such as the **Z and Score tests** or Wald test.

➡ Other Qs to ask, e.g. Is the sample size large enough? How do we perform an exact test?

# Segregation Analysis - Autosomal Recessive Disease I

- ➡ **Uncertain genotype problem:** A specific mating type may not be selected on the basis of the phenotype of the parents:

Unaffected individuals can be  $dD$  or  $dd$

- ➡ **Sampling issue:** How to select families? What is the correct ascertainment procedure?

- ➡ **Example:** interested in the segregation ratio for mating type  $dD \times dD$  (predicted to have  $p = 1/4$  under the autosomal recessive model).

- ◆ For a pair of unaffected parents, three possible mating types:

$dd \times dd$ ,  $Dd \times dd$  or  $Dd \times Dd$

- ◆ Propose: select families (both parents unaffected) with at least one affected offspring.
- ◆ Rationale:  $dd \times dd$  or  $Dd \times dd$  mating types do not produce affected offsprings. Thus **only  $Dd \times Dd$  mating type will be selected!**

# Segregation Analysis - Autosomal Recessive Disease II

## ► Problems of the above ascertainment procedure

- ◆ Will all matings with the  $Dd \times Dd$  type be randomly selected?
- ◆ An offspring of  $dD \times dD$  mating type has probability of 1/4 being affected.
- ◆ So, such sampling procedure may miss those  $dD \times dD$  families that have no affected offspring just by chance (also depending on the size of a family).
- ◆ Therefore the proportion of affected tends to be overestimated based on this sampling scheme.
- ◆ e.g. consider an extreme case where all families have only one child. If we require at least one affected offspring, then all offsprings in the selected sample will be affected!

## ► Statistical remedy: need to take into account of the “incomplete selection” of a mating type in segregation analysis. **Ascertainment procedure** should be clearly defined and accounted for (advanced stat gene topic).

- ◆ Glidden and Liang (2002). [Ascertainment adjustment in complex diseases](#). Genetic Epidemiology.
- ◆ Comments by Epstein (209-213), by Burton (214-218), and by Glidden (219-220) in the same issue of Genetic Epidemiology.

# Segregation Analysis - Beyond the Simple Model I

## ➡ Interpretation of deviation from Mendelian segregation ratios.

- ◆ More than one causal locus.
- ◆ Incomplete penetrance.
- ◆ Other characteristics of complex traits/diseases such as heterogeneity, environmental effect and gene-environment interactions.

## ➡ Advanced research in segregation analysis.

- ◆ Examine more complex and realistic models: allowing for additional genetic and non-genetic covariates, gene-environment interaction, and simultaneous consideration of linked markers, etc.
- ◆ E.g. In the likelihood formulation, let  $p$  the allele frequency and  $f_0, f_1, f_2$  the penetrance probabilities to be the parameters of interest; write the likelihood for the phenotypes  $Y$  as a function of these parameters; then calculate MLE.
- ◆ Segregation analysis for quantitative traits.

# Exercises

- ➡ Chapter 4 Exercise 1.
- ➡ Chapter 4 Exercise 2.
- ➡ Chapter 4 Exercise 3.
- ➡ Chapter 4 Exercise 7.
- ➡ Chapter 4 Exercise 11.
- ➡ Chapter 4 Exercise 13.

