Module 8 - Basic Concepts of (Full-Parametric) Linkage Analysis

(Fundamentals of) Statistical Genetics

Lei Sun

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

Outline

Chapter 6 - Basic Concepts of (Full-Parametric) Linkage Analysis

- Building map: marker-marker linkage analysis.
- Mapping gene: marker-trait linkage analysis
- Linkage analysis - one family and recombination status observed
- Linkage analysis - combining multiple families
- LOD Score and Function
- Genome-wide linkage analysis
- Linkage analysis - recombination status unobserved
- Multipoint linkage analysis
- General framework for full parametric linkage analysis
- Limitations and need for alternative linkage method.

Building Map and Linkage Analysis

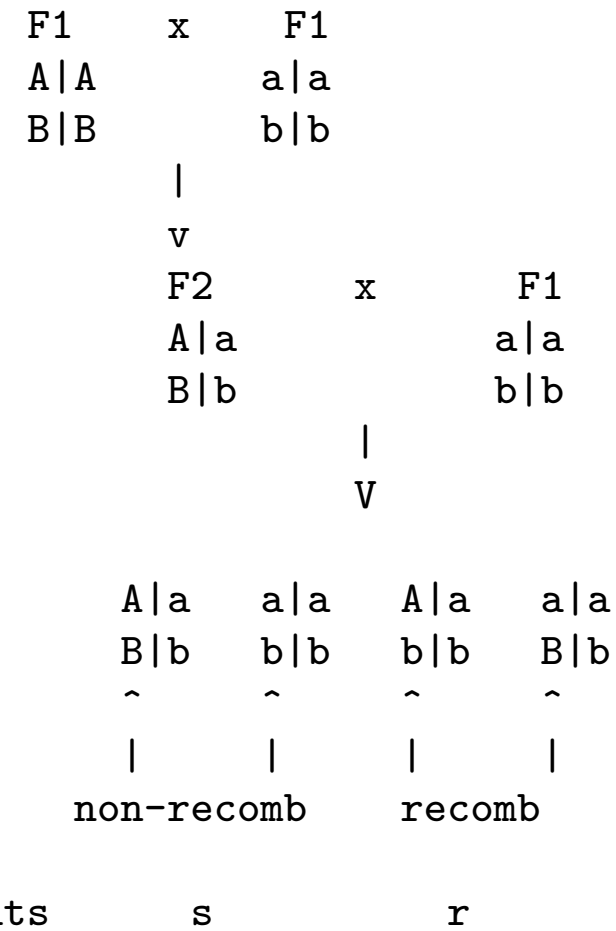
- Map is to order genetic markers and provide distance, $\theta_{(\text{marker 1, marker 2})}$, between markers, thus set up benchmarks along the genome.
- **Building map is essentially is a marker-marker linkage analysis**!

- Once we have the map, we can try to localize the DSL, i.e. find the marker/benchmark 'close', $\theta_{(\text{marker*, D})}$, to the unknown gene that influences the trait of interest .
- **Gene mapping study is essentially marker-trait linkage analysis.**

Building Map - Simple Case I

Simple Case: recombination status may be determined unambiguously.

- Because markers are highly polymorphic, or
- Use special experimental design e.g. the backcross design in animal model:



- Two different strains of animal.
- Inbred for > 40 generations (brother-sister matings).
- Obtain F_1 generation: homozygous across the whole genome.
- Obtain F_2 generation and backcross mating with F_1 generation.

Building Map - Simple Case II

- **Estimation of the distance**, again we should recognize that $r \sim Bino(n, \theta)$.

$$\hat{\theta} = \frac{r}{r+s} = \frac{r}{n}.$$

  - $\hat{\theta} \approx 1/2$: independent assortment (unlinked),
  - $\hat{\theta} \approx 0$: completely linked.

- **Hypothesis test**, $H_0 : \theta = 1/2$ (i.e. no linkage), e.g the Pearson χ^2 test:

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(r - n/2)^2}{n/2} + \frac{(s - n/2)^2}{n/2}$$

$$= \frac{2r^2 - 2n(r+s) + 2s^2 + n^2}{n} = \frac{2r^2 + 2s^2 - n^2}{n}$$

$$= \frac{2r^2 + 2s^2 - (r+s)^2}{n} = \frac{(r-s)^2}{n} \sim \chi^2_1.$$

Building Map - Simple Case III

- Recall Morgan's fruit flies study:

```
        C|c     x   c|c
        D|d         d|d
        red         purple
        normal      vestigial
                |
                V

      C|c     C|c       c|c     c|c
      D|d     d|d       D|d     d|d
      red     red       purple  purple
      normal  vestigial normal  vestigial

       s       r         r       s
counts 1339    151       154     1195
```

Building Map - Simple Case IV

- The distance between the two genes that determine the eye colour and wing length can be estimated by

$$\hat{\theta} = \frac{151 + 154}{151 + 154 + 1339 + 1195} = \frac{305}{2839} \approx 0.107$$

- Genetic distance based on the Haldane map function

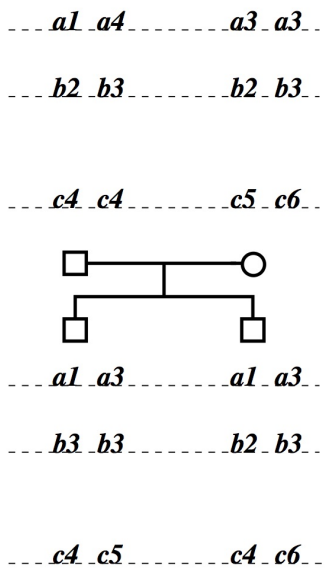
$$-log(1 - 2\theta)/2 \approx 0.12 \text{ Morgan, or } 12\text{cM}.$$

- Do we have enough evidence to reject $H_0 : \theta = 1/2$?

- Also check the example shown in Figure 6.2 of the Textbook ($r = 2$ and $n = 5$).

Building Map - Complex Case I

Complex Case: recombination status cannot be observed, because markers are not informative, or experimental design is not appropriate (human data).



Building Map - Complex Case II

- Direct counting method or Pearson χ^2_r test does not work in this case.
- **Likelihood approach more applicable**: give a statistical measure, based on the likelihood, for any possible order and associated set of distances.
- Again, we need to know how to write down likelihood over general pedigrees.

- Even so, the inference can be complex: n markers has $n!/2$ unique orders (reserved orders are equivalent likelihood-wise).
- If we calculate the likelihood for each order, and choose the one that maximizes the likelihood.

		time needed for analyses	
n	#orders	if 1/second	if 5/second
8	20160	5.6 hours	1.1 hours
12	239,500,800	7.6 years	1.5 years

- Need a heuristic and automated approach to reduce the number of orders to be examined.

Weeks (1991). Human linkage analysis: strategies for locus ordering. In: Advanced techniques in chromosome research. edited by K.W. Adolph. New York: Marcel Dekker. pp.297-330.

On Statistical Understanding and Use of Map

- Each map has unique characteristics (experimental techniques, methods of data collection, data quality, etc.).

- Statistical confidence on the ordering.
- Accuracy of distance estimations.

- Limits of resolution.
- Most errors cannot be easily detected and are not flagged.

- To minimize the effect of map errors:
  - Understanding map limitations.
  - Compare and utilize multiple maps.
  - Verify critical marker placements.
  - **Improve the robustness of down-stream methods that use maps**.

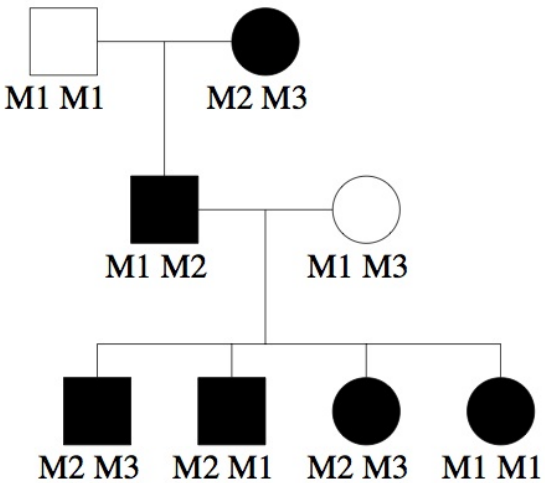
Building Map vs. Mapping Gene

- Similarity
  - Building map is essentially is a **marker-marker linkage analysis**!
  - Mapping gene is essentially is a **marker-trait linkage analysis**!

- Difference
  - In marker-marker linkage analysis: **genotype data of markers are known** (haplotype/phase/recombination may not be known as discussed before).
  - In marker-trait linkage analysis: **genotype data of the gene are NOT available** (inference would be even more complex).

Linkage Analysis - A Simple Example I

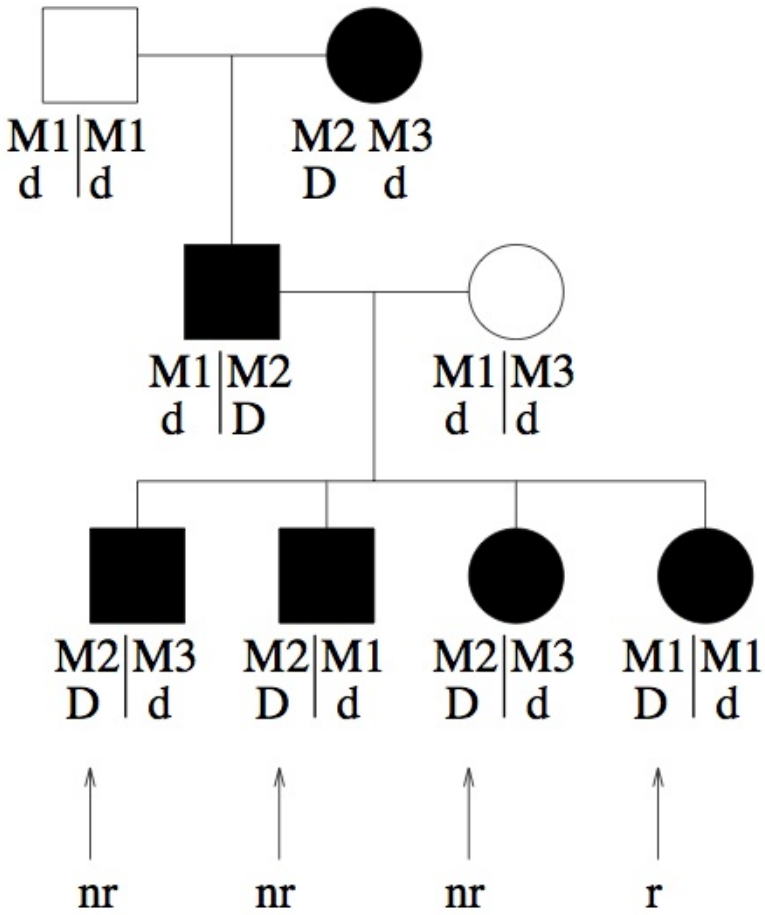
- Given a trait of interest,
  - Familial Aggregation $\rightarrow$ trait has genetic component.
  - Segregation analysis $\rightarrow$ mode of inheritance, e.g. autosomal dominant
    affected: DD or $Dd(dD)$; unaffected: dd.
- Given a genetic marker M:
  - The map position on the chromosome is known.
  - Genotype data at the marker can be collected, e.g.



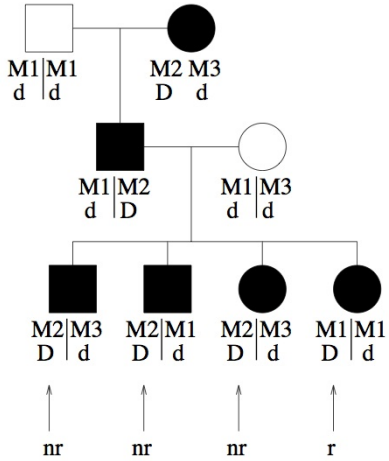
- Q: is the gene affecting the trait close/in linkage to this marker M or not?

Linkage Analysis - A Simple Example II

- Based on the assumed disease model, we can infer the genotype data at the gene locus! (Recall that for dominant disease, affected are assumed to have genotype $Dd(dD)$ rather DD because p_D is assumed to be very small.)



Linkage Analysis - A Simple Example III



- Note that the genotype of the affected grandmother is not phased.
- Genotypes of the two parents are both phased, but recombination status cannot be determined.
- Genotypes of the four grandchildren can all be phased, but recombination status can only be determined for the chromosome inherited from the father. (Because mother is homozygous *dd* then we don't know the origin of the transmitted *d*.)

- So we have $n = 4$ and $r = 1$. What is the inference of θ, the distance as measured by recombination fraction between the known M locus and the unknown D locus?

Linkage Analysis - A Simple Example IV

- Again we use binomial distribution to model the count data:

$$r \sim Bino(n, \theta).$$

- Note that θ is bounded between 0 and $1/2$.
- (Kernel of the) Likelihood

$$L(\theta) = \theta^r (1-\theta)^{(n-r)}.$$

- Loglikelihood

$$l(\theta) = rlog(\theta) + (n-r)log(1-\theta).$$

- Score function

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \frac{r}{\theta} - \frac{n-r}{1-\theta} = \frac{r - n\theta}{\theta(1-\theta)}.$$

- MLE

$$S(\theta) = 0 \Longrightarrow \hat{\theta} = \frac{r}{n} = \frac{1}{4}.$$

- So the recommendation fraction is estimated to be 0.25 which is less than the null of no linkage of 0.5, but are we convinced that the DSL is linked to M?

Linkage Analysis - A Simple Example V

- Testing the null hypothesis of no linkage

$$H_0 : \theta = \frac{1}{2}.$$

- The LRT statistic

$$T = 2log\frac{L(\hat{\theta})}{L(\theta_0)} = 2log\frac{\hat{\theta}^r(1-\hat{\theta})^{n-r}}{\frac{1}{2}^n}$$

- What is the distribution of T under H_0?

- Note that $\hat{\theta}$ is restricted to be $\leq \frac{1}{2}$, that is (e.g. $r = 3$ and $n = 4$)

$$\hat{\theta} \equiv \frac{1}{2}, \text{ if } \frac{r}{n} > \frac{1}{2}$$

- Also note that under H_0, $\frac{r}{n} > \frac{1}{2}$ with probability $\frac{1}{2}$ (exactly if n was an odd number and approximately if n was even).

Linkage Analysis - A Simple Example VI

- The consequence of this is that the resulting test statistic, under the null hypothesis of no linkage H_0, T is χ^2_1 with probability $\frac{1}{2}$ and $= 0$ with probability $\frac{1}{2}$. That is the distribution of T is a mixture of $0.5 \cdot \chi^2$ and 0.5 mass in zero. Often denoted as

$$\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_0.$$

- Self and Liang 1987, Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions.

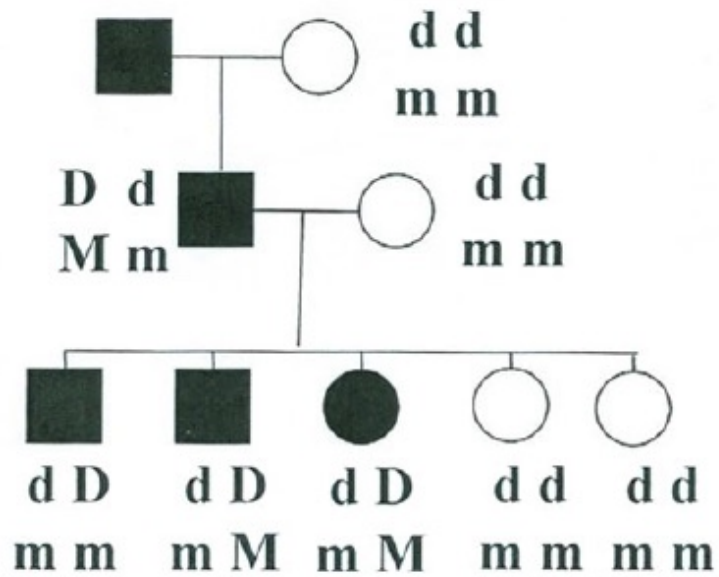
$$T_{obs} = 1.046 \Longrightarrow$$
$$\text{p-value } = P(T \geq T_{obs}) = 0.5 \cdot P(\chi^2_1 \geq T_{obs}) + 0.5 \cdot P(\chi^2_0 \geq T_{obs})$$
$$= 0.5 \cdot 0.306 + 0.5 \cdot 0 = 0.153.$$

- Are we surprised that, although $\hat{\theta} = \frac{1}{4} < \frac{1}{2}$, we cannot really reject the null that DSL is NOT linkage to M?

- Why the previous Pearson χ^2 test did not have the mixture distribution as discussed here?

Linkage Analysis - Another Simple Example

- Consider the family shown in Figure 6.4 of the Textbook.



- In this case, $r = 1$ and $n = 5$.

$$\hat{\theta} = \frac{r}{n} = \frac{1}{5}.$$

$$T_{obs} = 1.927 \Longrightarrow \text{ p-value } = 0.083.$$

Linkage Analysis - Multiple Families I

- How do we combine the data/evidence across the two families?
- If we assume families are independent of each other, then

$$r_1 \sim Bino(n_1, \theta) \perp r_2 \sim Bino(n_2, \theta)$$

$$L(\theta) = Prob(r_1, r_2; \theta) = Prob(r_1; \theta) Prob(r_2; \theta) = L_1(\theta) L_2(\theta).$$

- It will be additive on the log likelihood scale,

$$l(\theta) = log(L(\theta)) = log(L_1(\theta) L_2(\theta)) = l_1(\theta) + l_2(\theta).$$

- Note that the independence assumption can be false if there are cryptic relatedness between families; Identifying cryptic relationships.

Linkage Analysis - Multiple Families II

- What is the MLE based on the combined data?

$$\hat{\theta} = \frac{r_1 + r_2}{n_1 + n_2} = \frac{1+1}{4+5} = \frac{2}{9}.$$

- The above can be obtained by recognizing that the kernel of the joint likelihood has the same form as that of $r_1 + r_2 \sim Bino(n_1 + n_2, \theta)$ (though note that $P(r_1, r_2) \neq P(r_1 + r_2)$ because $\binom{n_1}{r_1}\binom{n_2}{r_2} \neq \binom{n_1+n_2}{r_1+r_2}$):

$$\begin{aligned} L(\theta) &= \theta^{r_1}(1-\theta)^{n_1-r_1} \cdot \theta^{r_2}(1-\theta)^{n_2-r_2} \\ &= \theta^{r_1+r_2}(1-\theta)^{n_1+n_2-(r_1+r_2)}. \end{aligned}$$

- What is the linkage evidence based on the combined data?

$$T_{obs} = 2.942 \Longrightarrow \text{ p-value } = 0.043.$$

Linkage Analysis - Multiple Families III

- Note that

$$T = 2log\frac{L(\hat{\theta})}{L(\theta_0)} = 2log\left(\frac{L_1(\hat{\theta})}{L_1(\theta_0)} \cdot \frac{L_2(\hat{\theta})}{L_2(\theta_0)}\right) = T_1 + T_2.$$

- However, we also notice that

$$T_{obs} = 2.942 \neq 1.046 + 1.927 = T_{obs,1} + T_{obs,2}.$$

- It is important to note that $T_{obs,1} = 1.046$ was calculated using $\hat{\theta}_1 = \frac{1}{4}$, the MLE estimated using only family 1 ($T_{obs,2} = 1.927$ using $\hat{\theta}_2 = \frac{1}{5}$ from family 2 only), so

$$T_{obs}(\hat{\theta}) \neq T_{obs,1}(\hat{\theta}_1) + T_{obs,2}(\hat{\theta}_2).$$

Linkage Analysis - Multiple Families IV

- But, if we use $\hat{\theta} = \frac{2}{9}$ the MLE estimated from the combined data, then we do have additivity across independent families.

$$T_{obs}(\hat{\theta}) = 2.942 = 1.029 + 1.913 = T_{obs,1}(\hat{\theta}) + T_{obs,2}(\hat{\theta}).$$

LOD Score

- In linkage analysis, the tradition is to use log with base 10 on the LR, but it has a simple connection with the LRT statistic T that we have discussed (Chapter 6, Exercise 2).

$$LRT = T = 2log\frac{L(\hat{\theta})}{L(\theta_0)}$$

$$LOD = log_{10}\frac{L(\hat{\theta})}{L(\theta_0)} = \frac{log_{10}e}{2} \cdot T \approx 0.217T$$
$$(T \approx 4.6LOD)$$

- *One reason for choosing base 10 instead of e in the logarithm is that it facilities interpretation. A lod-score of 1 says the $P(data|\theta)$ is 10 times $P(data|\theta = \frac{1}{2})$, and for a lod-score of 2, the ratio is 100, etc.*

- The additivity holds, as long as we plug in the MLE $\hat{\theta}$ estimated globally across the families.

$$LOD(\hat{\theta}) = \sum_{i \in families} LOD_i(\hat{\theta}).$$

LOD Function - Graphic Display I

- Instead of obtaining LR and LOD score at the MLE $\hat{\theta}$ value, we can sketch and study the **entire likelihood**, LR or LOD function w.r.t. θ. This function carries more information than a single MLE value.
- Consider the two families we have discussed so far, $r_1 = 1, n_1 = 4$ and $r_2 = 1, n_2 = 5$.
- The individual LOD functions (essentially scaled LR functions) w.r.t to θ are

$$LOD_i(\theta) = log_{10}\frac{L_i(\theta)}{L_i(\theta_0)} = log_{10}\frac{\theta^{r_i}(1-\theta)^{n_i-r_i}}{\frac{1}{2}^{n_i}}, \quad i = 1, 2$$

- The overall LOD function combining information across the two families

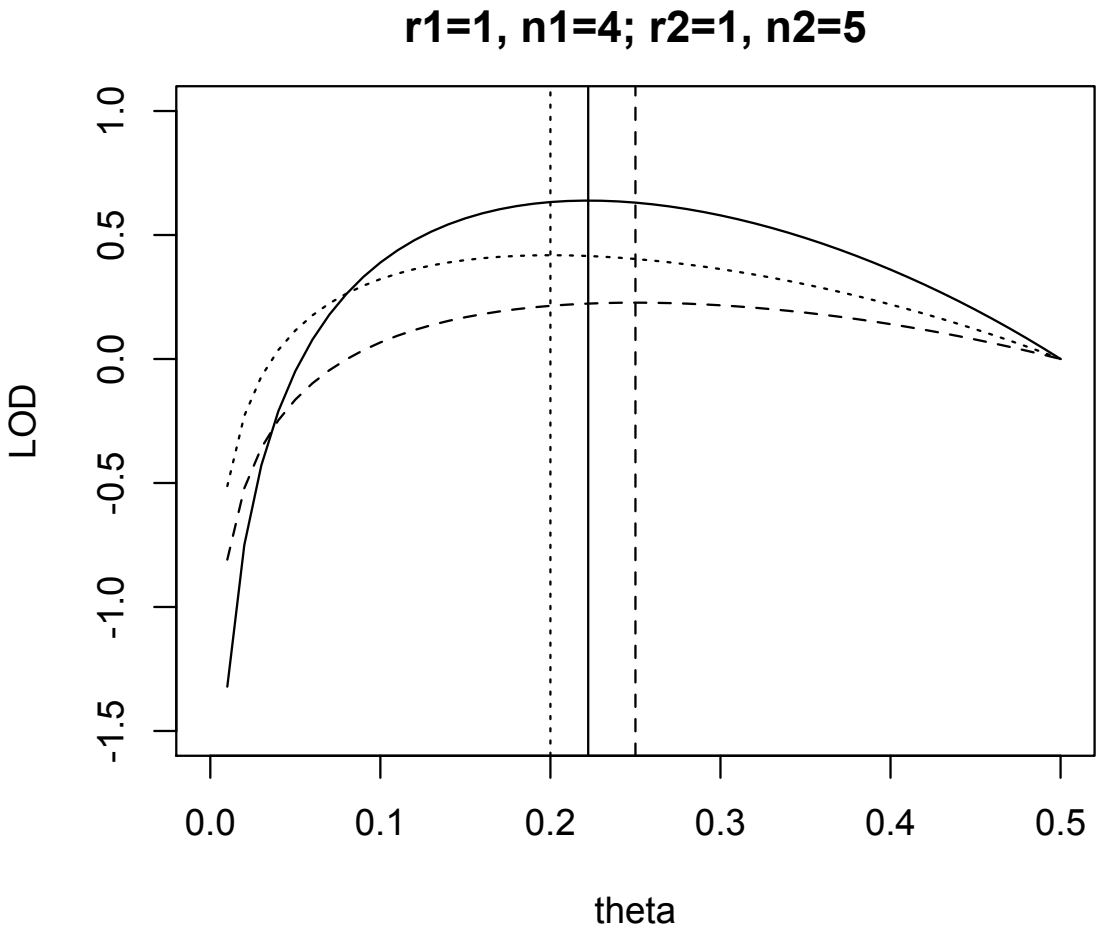
$$LOD_{combined}(\theta) = log_{10}\frac{L(\theta)}{L(\theta_0)} = log_{10}\frac{\theta^{r_1+r_2}(1-\theta)^{n_1+n_2-(r_1+r_2)}}{\frac{1}{2}^{n_1+n_2}}$$

$$= log_{10}\frac{L_1(\theta)\cdot L_2(\theta)}{L_1(\theta_0)\cdot L_2(\theta_0)} = log_{10}\frac{L_1(\theta)}{L_1(\theta_0)} + log_{10}\frac{L_2(\theta)}{L_2(\theta_0)} = LOD_1(\theta) + LOD_2(\theta).$$

- Note that the additivity holds as long as we plug in the SAME θ.

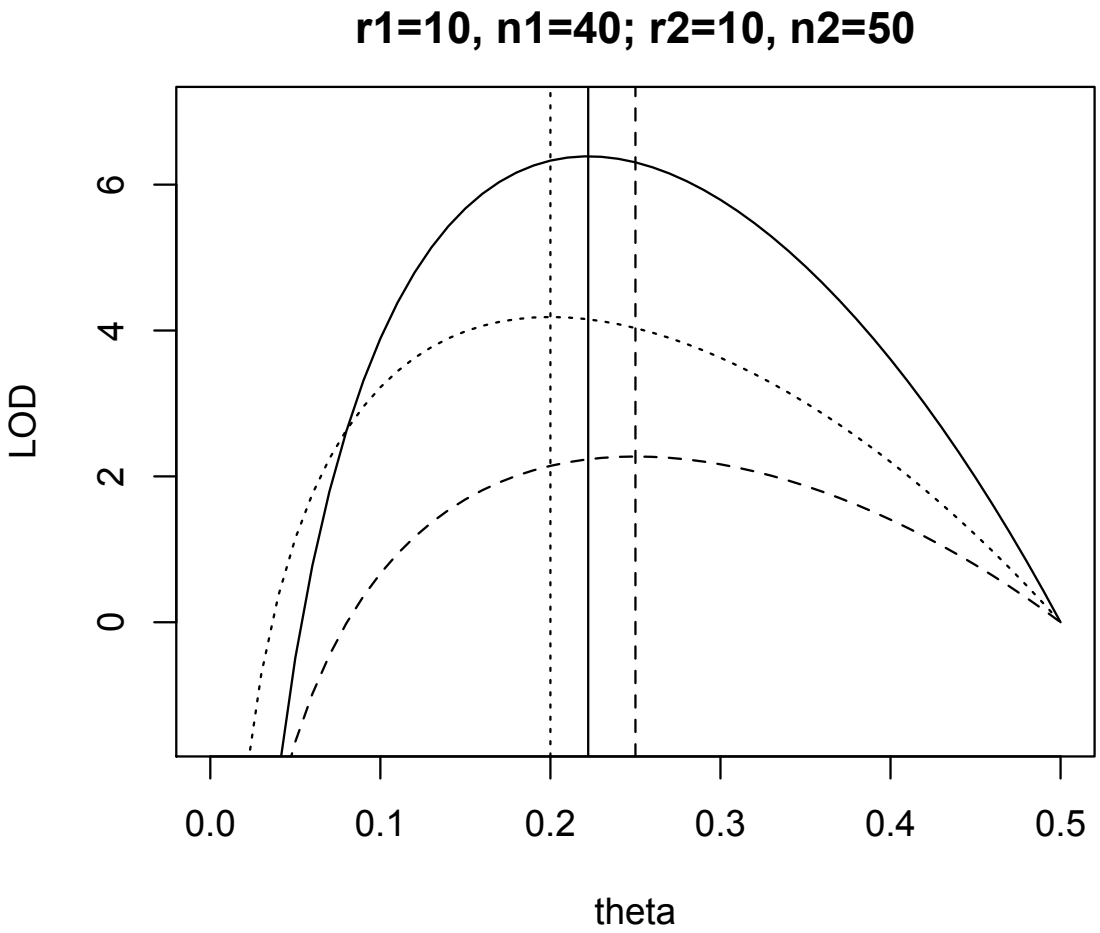
LOD Function - Graphic Display II

- R codes



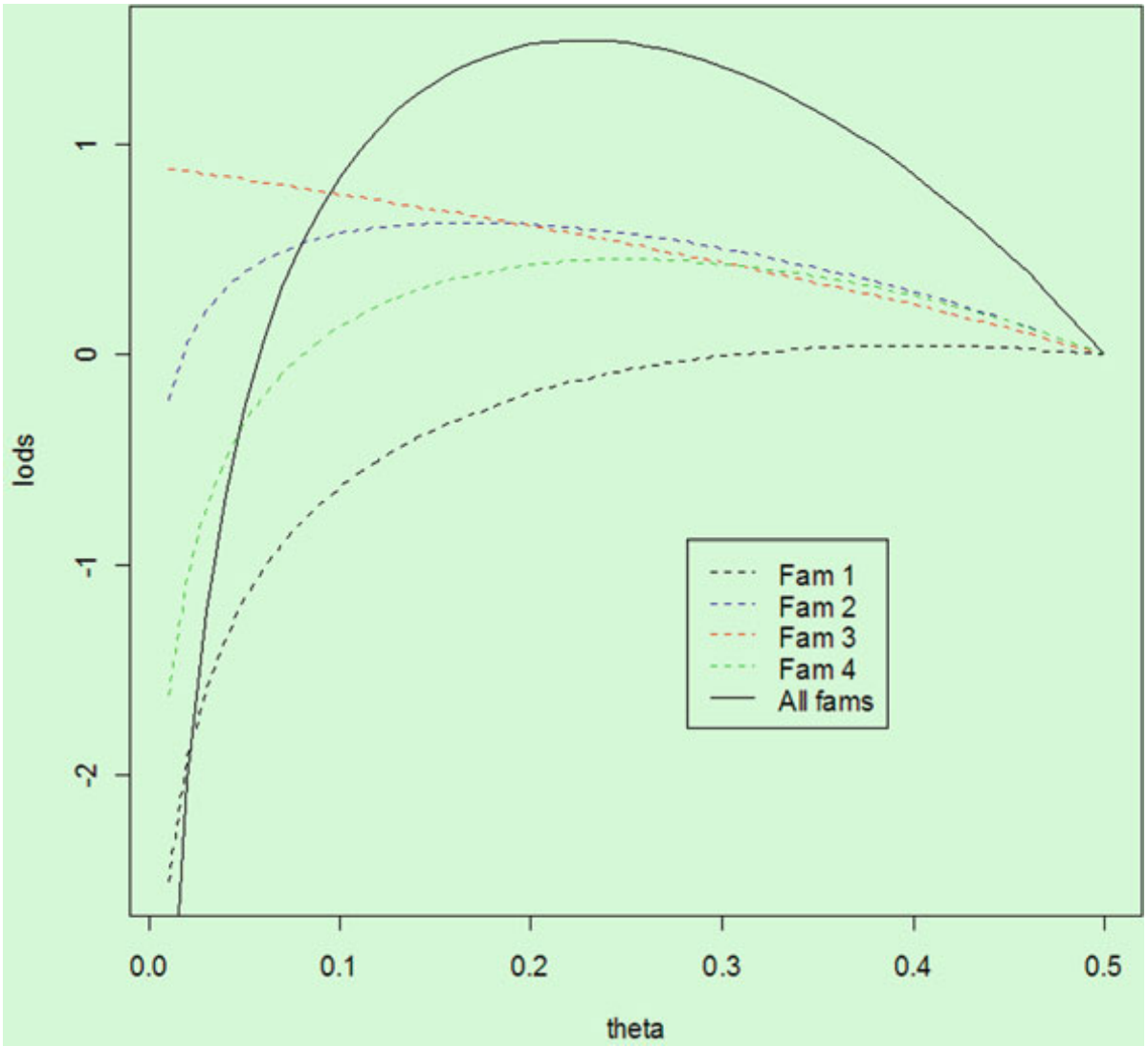
LOD Function - Graphic Display III

- What happen if we increased the sample size but keep the MLE the same?



LOD Function - Graphic Display IV

- Another example: Figure 6.3 of the Textbook involving four different families.



Genome-Wide Linkage Analysis and Significance Level I

- Location of the DSL for the phenotype of interest could be anywhere on the genome. So we need to do a whole-genome scan or genome-wide mapping studies.
- But how many markers/benchmarks do we need to place on the genome to perform a genome-wide linkage analysis?
- In theory, 23 markers because two markers are linked as long as they are on the same chromosome, and DSL must reside on one of the chromosomes.
  - e.g. consider the longest chromosome 1 which is about 300cM or 3cM, and if we place the marker in the middle of the chromosome,

$$\theta_{\text{marker, DSL}} = \frac{1 - e^{-2t}}{2} \leq \frac{1 - e^{-2 \cdot \frac{3}{2}}}{2} = 0.475.$$

  - If we have enough data, then we will be able to reject the null of no linkage, $H_0 : \theta = 0.5$.

Genome-Wide Linkage Analysis and Significance Level II

- BUT, even we reject the null, the DLS can be up to 150cM away from the linked marker, which is not very helpful! (Recall 1cM $\approx$ 1 million bp.)

- Instead, a standard whole genome linkage scan typically requires $\approx$ 500-1000 markers, placed $\approx$ 5cM-10cM apart across the genome.

$$\theta_{\text{marker, DSL}} = \frac{1 - e^{-2t}}{2} \leq \frac{1 - e^{-2 \cdot \frac{0.1}{2}}}{2} = 0.048.$$

- This is much better than before, but 5cM $\approx$ 5 million bp, so we still need additional methods to determine a more precise location for the DLS. Hence the LD-based fine-mapping or association studies which we discuss later.

Genome-Wide Linkage Analysis and Significance Level III

- One remaining question: how small the p-value should be for testing $H_0 : \theta = \frac{1}{2}$ for each of the 1000 or so markers?
- Why LOD score is typically required to be > 3.3 to declare significant linkage evidence (related to Chapter 6, Exercise 2).

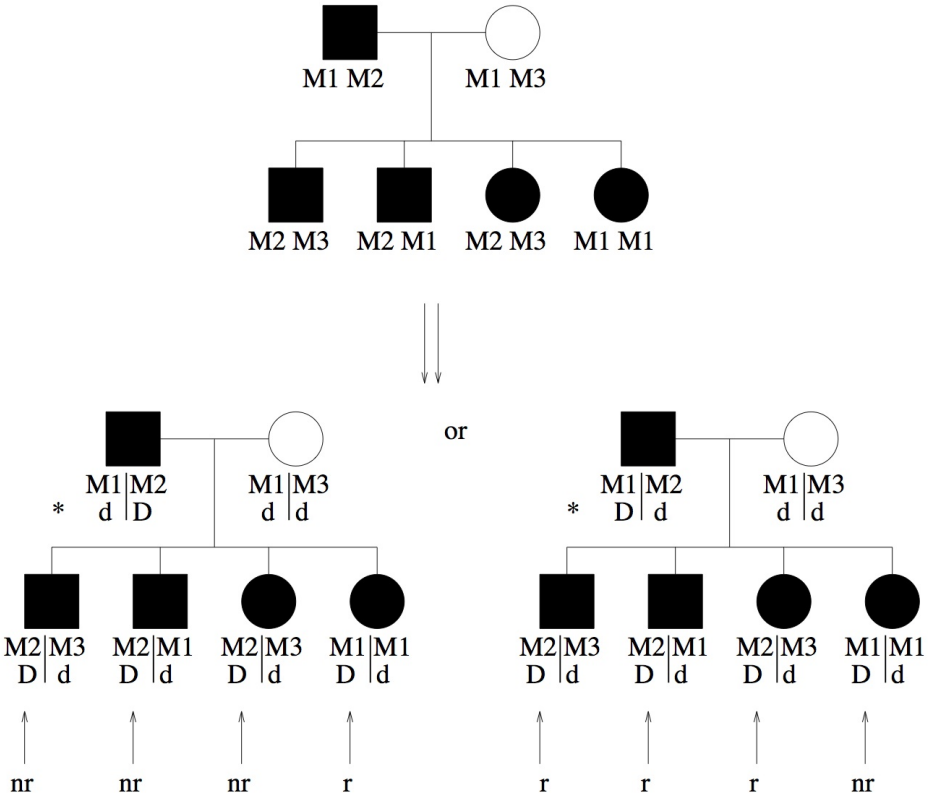
$$T_{obs} \approx 4.6 LOD_{obs} = 4.6 \cdot 3.3 = 15.18.$$

$$\text{p-value } = 0.5 \cdot P(\chi^2_1 > T_{obs}) \approx 0.00005 = 5x10^{-5}.$$

- Related to multiple hypothesis testing which we discuss later.
- If we used the traditional $\alpha = 0.05$ type 1 error threshold, then we would expect $0.05 * 1000 = 50$ markers on average to be declared significantly linked with the DLS even if the trait of interest had no genetic causes (i.e. no DLS in the first place)! If we used $5x10^{-5}$?

Linkage Analysis - A More Complex Example I

- In the Example 1, what would happen if there was no information available for the grandparents?



Linkage Analysis - A More Complex Example II

- Both haplotype configurations are equally likely (need to assume no LD, a reasonable assumption for linkage data: the unknown DLS location is unlikely to be extremely close to any of the markers placed on the genome.)

$$L(\theta) = \frac{1}{2} \cdot \binom{4}{1} \theta^1 (1-\theta)^3 + \frac{1}{2} \cdot \binom{4}{3} \theta^3 (1-\theta)^1.$$

- Only need to consider kernel of the likelihood:

$$L(\theta) = \theta^1 (1-\theta)^3 + \theta^3 (1-\theta)^1.$$

Linkage Analysis - A More Complex Example III

- What would be the MLE now? Intuitively should be $\hat{\theta} = \frac{1}{2}$, and formally:

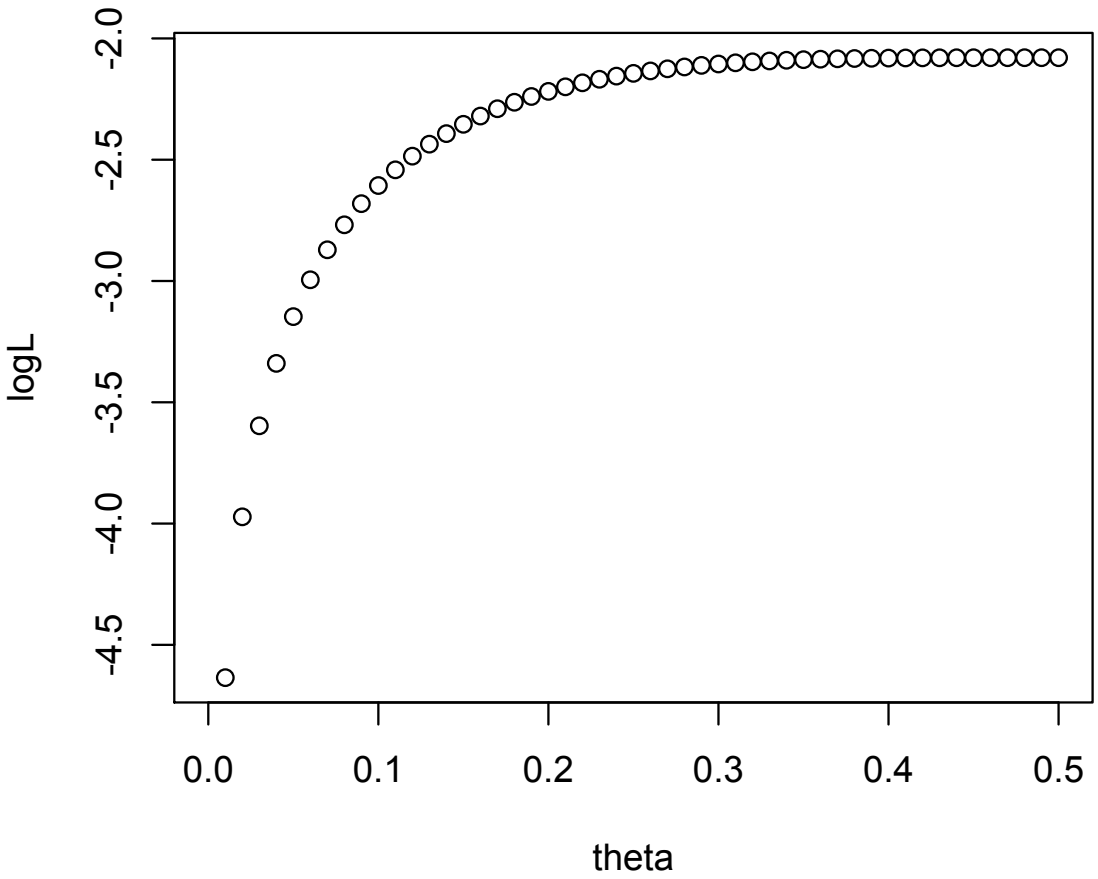
$$l(\theta) = log\left(\theta^1(1-\theta)^3 + \theta^3(1-\theta)^1\right).$$

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \frac{(1-\theta)^3 - \theta 3(1-\theta)^2 + 3\theta^2(1-\theta) - \theta^3}{\theta^1(1-\theta)^3 + \theta^3(1-\theta)^1}.$$

$$S(\theta) = 0 \Longrightarrow \hat{\theta} = \frac{1}{2}.$$

Linkage Analysis - A More Complex Example IV

- We can also look the log likelihood function. R codes



Linkage Analysis - A More Complex Example V

- There is no point of testing $H_0 : \theta = \frac{1}{2}$, but if we went ahead anyway:

$$L(\theta) = \theta^1(1-\theta)^3 + \theta^3(1-\theta)^1.$$

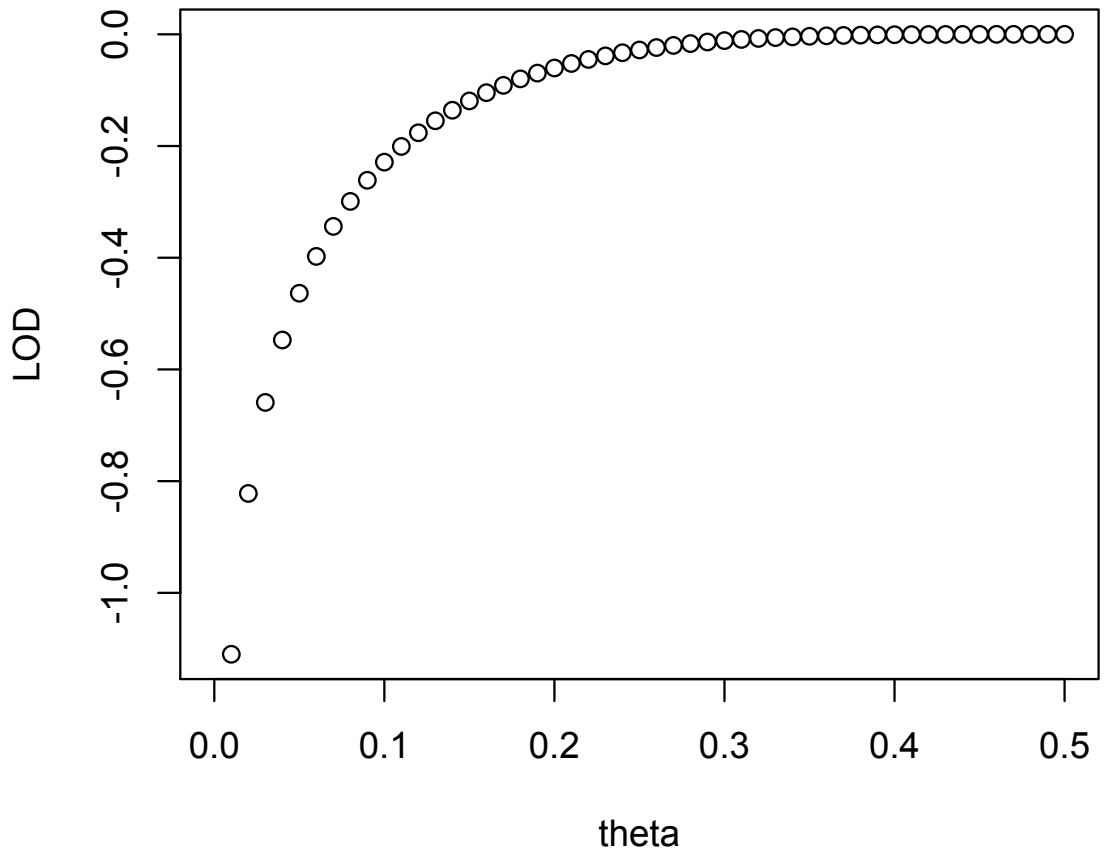
$$T_{obs} = 2log\frac{L(\hat{\theta})}{L(\theta_0)} = 2log\frac{\hat{\theta}^1(1-\hat{\theta})^3 + \hat{\theta}^3(1-\hat{\theta})^1}{\frac{1}{2}^4 + \frac{1}{2}^4} = 2log(1) = 0.$$

- $LOD_{obs} = 0$; how would the LOD function look w.r.t θ?

$$LOD(\theta) = log_{10}\frac{L(\theta)}{L(\theta_0)} = log_{10}\frac{\theta^1(1-\theta)^3 + \theta^3(1-\theta)^1}{\frac{1}{2}^4 + \frac{1}{2}^4}.$$

Linkage Analysis - A More Complex Example VI

- Not surprisingly the maximum LOD is zero and achieved at $\theta = \frac{1}{2}$.



Linkage Analysis - Another More Complex Example I

- aa
- Now consider Example 2 (Figure 6.4 of the Textbook), *If the grandmother's marker data is missing, we have no way of determining phase in the father*, and we will also have a similar (kernel) likelihood.

$$L(\theta) = \theta^1(1-\theta)^4 + \theta^4(1-\theta)^1.$$

- What is the MLE?

$$l(\theta) = log\left(\theta^1(1-\theta)^4 + \theta^4(1-\theta)^1\right).$$

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \frac{(1-\theta)^4 - \theta 4(1-\theta)^3 + 4\theta^3(1-\theta) - \theta^4}{\theta^1(1-\theta)^4 + \theta^4(1-\theta)^1}.$$

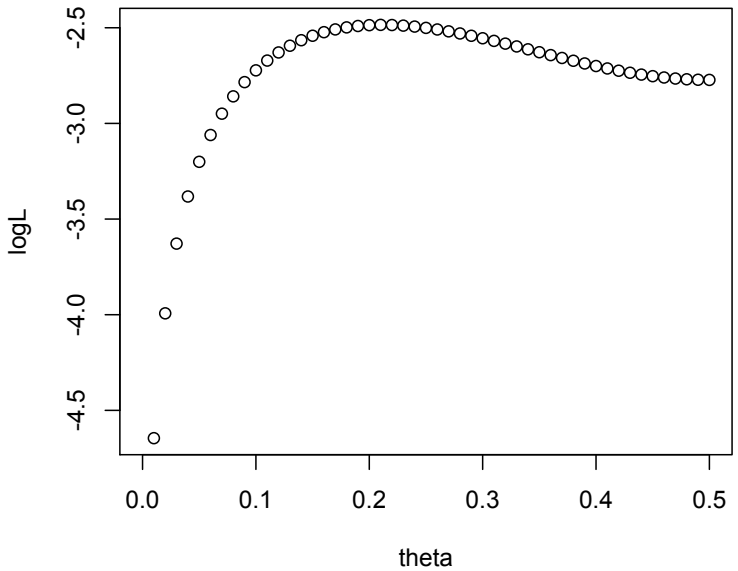
$$S(\theta) = 0 \Longrightarrow \hat{\theta} = ???$$

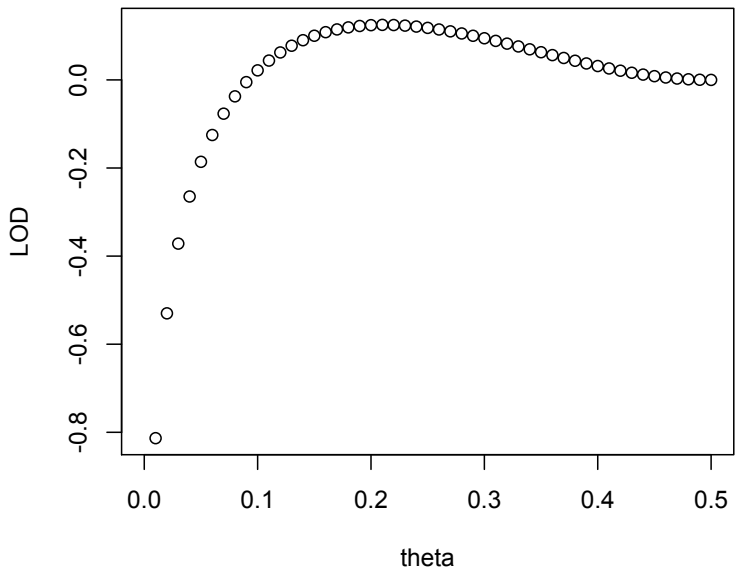
Linkage Analysis - Another More Complex Example II

$$LOD(\theta) = log_{10}\frac{L(\theta)}{L(\theta_0)} = log_{10}\frac{\theta^1(1-\theta)^4 + \theta^4(1-\theta)^1}{\frac{1}{2}^5 + \frac{1}{2}^5}.$$

Linkage Analysis - Another More Complex Example III

- Here we see an example that the solution to $S(\theta) = 0$ may not be unique! How do you check to make sure that the solution is the MLE? R codes





Linkage Analysis - Combining Complex Examples I

- Now if we combine the two families, then

$$L(\theta) = \left(\theta^1(1-\theta)^3 + \theta^3(1-\theta)^1\right) \cdot \left(\theta^1(1-\theta)^4 + \theta^4(1-\theta)^1\right).$$

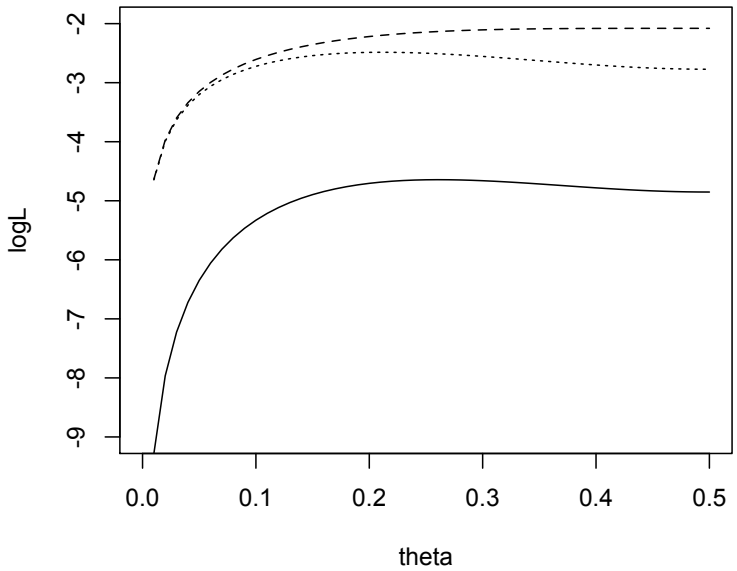
$$l(\theta) = log(\theta^1(1-\theta)^3+\theta^3(1-\theta)^1)+log(\theta^1(1-\theta)^4+\theta^4(1-\theta)^1) = l_1(\theta)+l_2(\theta).$$

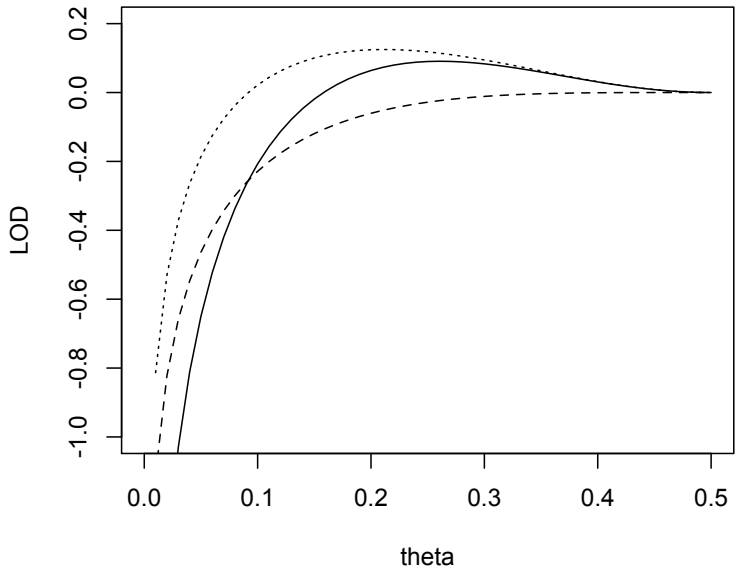
$$LOD(\theta) = LOD_1(\theta) + LOD_2(\theta).$$

- What would be the MLE?
- What would the LOD function look like?

Linkage Analysis - Combining Complex Examples II

- Are we surprised that the max LOD decreased after we combining these two families? R codes





- *Calculating lod-scores separately for each family is also useful in the presence of genetic heterogeneity, i.e., different mutations may be responsible for the same disorder in different families.*

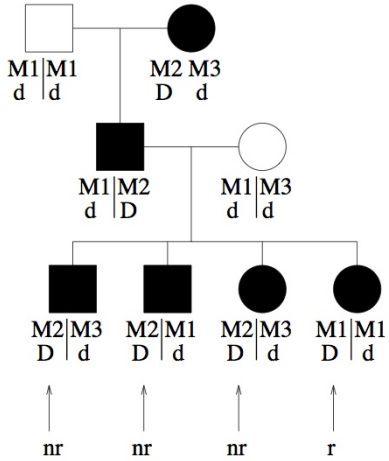
Multipoint Linkage Analysis I

- The linkage analysis we have learned so far focuses on inference of $\theta_{DSL,M}$ using the genotype information at this Marker only. This is also called **two-point linkage analysis**.

- **Multipoint linkage analysis** still focus on inference of $\theta_{DSL,M}$, but using the genotype information at all availalbe markers that on the same chromosome as the current Marker of interest.

- Markers on the same chromosome are linked to each, thus provide inheritance/halotype/recombination information to each!

Multipoint Linkage Analysis II



- Revisit Example 1. If marker N is also genotyped, consider the mid-generation mother with homozygous *dd* genotype, If we observe data

```
M1 M3
d  d
N1 N2
   |
   v
M3   M1   M3   M1
d    d    d    d
N2   N1   N2   N1
```

- Now can we say something about the recombination status for the haplotypes inherited from there mother?

Multipoint Linkage Analysis III

- This would depends on how close the two markers are. If the two markers are linked at $\theta = 0.05$, about 5 cM away, then most likely we can phase the mother's genotypes.

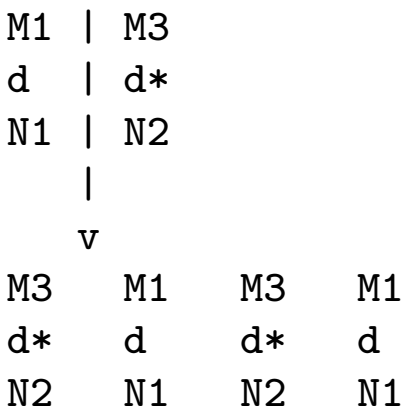
```
M1 | M3
d  | d
N1 | N2
   |
   v
M3   M1   M3   M1
d    d    d    d
N2   N1   N2   N1
```

- Because the other haplotype configuration would imply 4 recombinations happened at the offspring generation, with probability θ^4.

```
M1 | M3
d  | d
N2 | N1
   |
   v
M3   M1   M3   M1
d    d    d    d
N2   N1   N2   N1
```

Multipoint Linkage Analysis IV

- Now, if we look at the DSL, then we are likely to determine the origins of the inherited *d*s and infer that there are no-recombinations between marker M and DSL.



- Thus, when we infer $\theta_{DSL,M}$ (or later on about IBD sharing probabilities for non-parametric linkage analysis), multipoint linkage analysis will also consider the information provided at locus *N* (in fact all markers from the same chromosome).
- **Linked markers provide inheritance information to each other!**

Two-point vs. Multipoint Linkage Analysis

- Multipoint linkage analysis is more powerful than two-point linkage analysis.
- Multipoint linkage analysis can provide LOD score for each point on the chromosome, not just at the observed marker loci, because we can think of each point as a marker with missing data and use all the other linked markers to provide information for the linkage inference at that point.

  (Two-point linkage analysis extrapolates LOD score by connecting the LOD scores at two adjacent markers using a smooth curve.)

- However, multipoint analysis also has its pitfalls!
  - Marker map has to be known precisely.
  - If map is not accurate or even wrong, multipoint analysis may introduce more "noise" to the analysis.
  - The analysis is more computationally intensive even with recent improvements in algorithms.
  - There are different algorithms and software packages (e.g. GENEHUNTER) available for the calculation and assessment of likelihood.

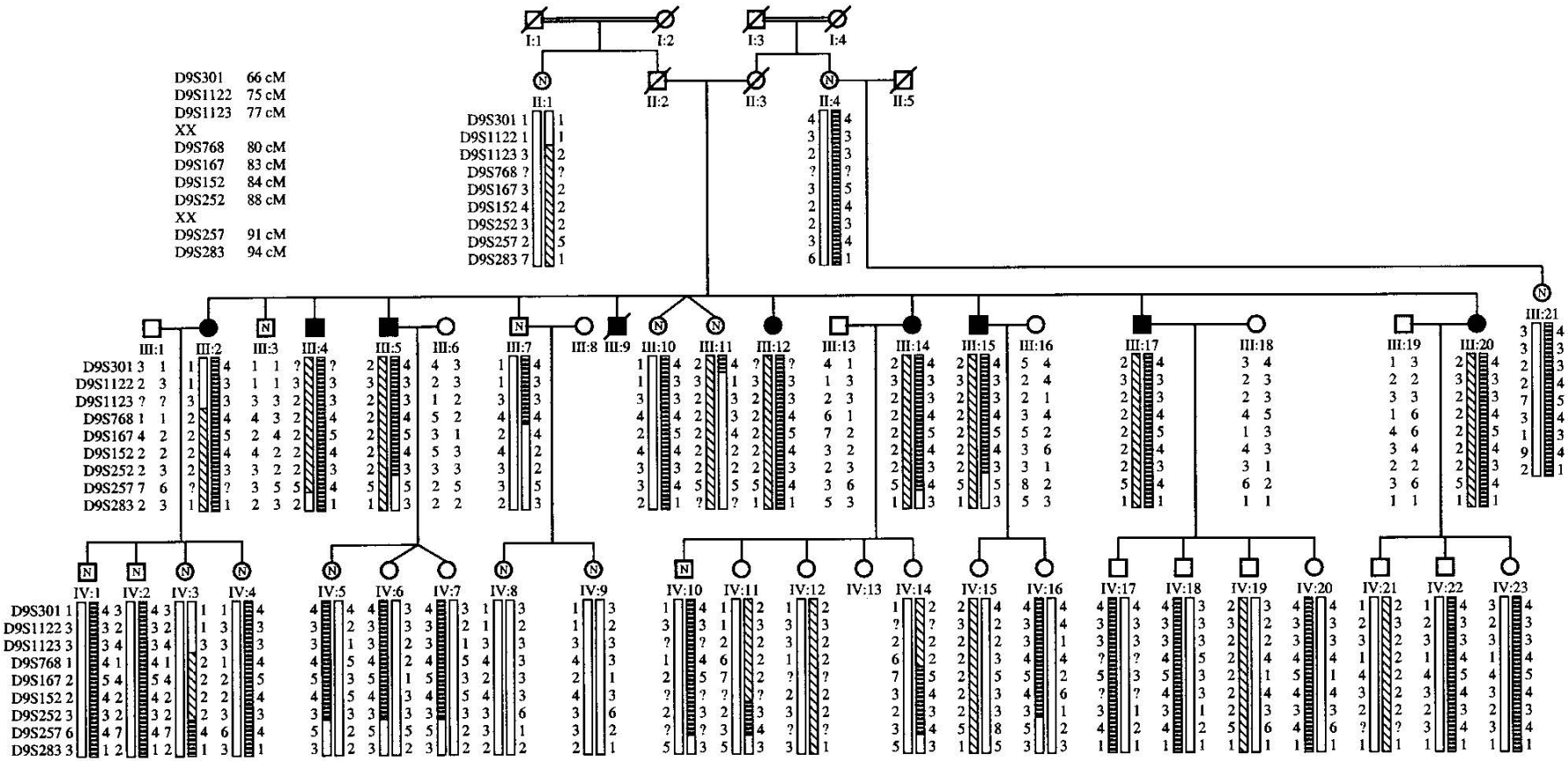
- In practice, conduct both two-point and multipoint linkage analyses and compare.

An Example of Parametric Linkage Analysis I

- Heon et al. (2001). A progressive autosomal recessive cataract locus maps to chromosome 9q13-q22. American Journal of Human Genetics 68:772-77.
- Cataracts (age-related) is a leading case of blindness in most countries.
- Proposed model:
  - Autosomal recessive (previous segregation analysis)
  - Complete penetrance and no phenocopy
  - Disease allele frequency $= 0.01$.
- Families collected: a large family with 9 affecteds (Figure 1).

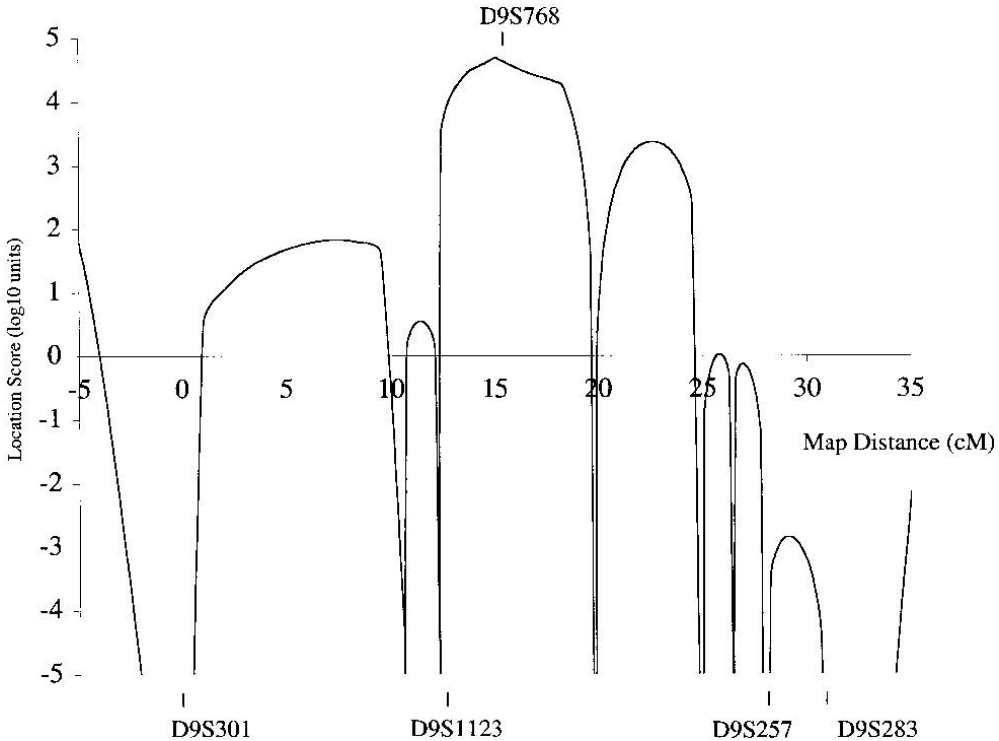
Genealogy and summarized haplotype showing the most informative markers. All spouses were examined and found to be normal. Blackened symbols indicate clinically affected individuals; unblackened symbols represent unaffected relatives. Unblackened symbols containing an "N" indicate relatives ≥ 30 *years old who were examined but who did not show signs of the disease; empty unblackened symbols indicate unaffected relatives who are* ≤ 30 *years old; slashes indicate deceased individuals. The hatched boxes indicate the affected haplotype, whereas the unhatched boxes indicate the unaffected haplotype. "XX" indicates the beginning and end of the disease-gene interval. "?" indicates the genotype was not available.*

An Example of Parametric Linkage Analysis II



An Example of Parametric Linkage Analysis III

- Genome-wide linkage analysis/scan
  - 380 microsatellite markers spaced at about 10 cM apart
  - Two-point linkage analysis using SIMWALK2.
  - Results: Table 3 and Figure 3 in the paper.
  - A maximum LOD score is about 4.7 at marker D9S768



An Example of Parametric Linkage Analysis IV

Table 3

Two-Point Linkage Data for ARPC Phenotype and Markers of the 9q13-q22 Region

Marker (Distance [cM])[a]	Location	Heterozygosity[b]	LOD Score at θ = 0	.1	.2	.3	.04	Z_{max}	θ_{max}
D9S301 (66)	9p21-9q21	.71	∞	1.20	1.13	.76	.27	1.23	.13
D9S1122 (75)	9pter-qter	.67	∞	2.90	2.32	1.52	.6	3.02	.05
D9S1123 (77)	9pter-qter	.66	∞	.96	.84	.56	.21	.96	.11
D9S153 (79)	9q13-q22.3	.68	2.00	1.63	1.23	.78	.29	2.00	.00
D9S1867 (79)	9pter-qter	.71	2.00	1.62	1.21	.75	.28	2.00	.00
D9S768 (80)	9q13-q22.3	.79	4.71	3.85	2.92	1.89	.76	4.71	.00
D9S167 (83)	9q13-q22.3	.83	2.30	1.94	1.49	.98	.40	2.30	.00
D9S152 (84)	9q21-q22	.72	2.00	1.38	.76	.24	.02	2.00	.00
D9S1119 (85)	9pter-qter	.64	.04	.02	.01	.00	.00	.14	.00
D9S252 (88)	9q13-q22	.66	2.00	1.38	.76	.24	.02	2.00	.00
D9S1812 (90)	9pter-qter	.57	2.30	1.93	1.48	.97	.40	2.30	.00
D9S257 (91)	9q13-q22	.84	∞	1.73	1.57	1.08	.42	1.74	.12
D9S283 (94)	9q13-q22	.75	∞	.48	.80	.67	.30	.80	.21

Linkage Analysis General Framework I

- Family Data, Y and R (Pedigree structure): families with individuals affected by the disease of interest. Typically large and multi-generation families.
- Genotype Data (G): a larger number of markers placed and genotyped throughout the genome.
- The analysis: infer θ between a known marker/position on the genome and the unknown DSL, D.
  - Estimate θ; obtain MLE estimate.
  - Test θ against the null of no linkage $H_0 : \theta = 0.5$; obtain LOD.
  - Perform the analysis for each marker or location on the genome; obtain Maximum LOD score and corresponding location.
  - Adjust for multiple hypothesis testing.

Linkage Analysis General Framework II

- Main assumptions
  - Assume that a particular parametric model is true. The model specifies the mode of inheritance for the trait under the study, and it explains the observed inheritance pattern within families, e.g.
  - p: frequency of the disease allele $P(D)$.
  - f_0, f_1, f_2: penetrance, $f_i = P(Y = 1|i \text{ copies of } D)$.

- In the examples discussed so far, we assumed complete penetrance with no phenocopy, so $f_i = 0$ or $f_i = 1$ and we can infer the genotypes at the D locus for each individual.
- However, in practice, such simple model is not realistic. $P(Dd|\text{affected})$ and $P(dd|\text{unaffected})$ are both part of the likelihood which depends on the values of the parameters.

$$P(Dd|affected) = \frac{P(Dd, affected)}{P(affected)} = \frac{f_1 2p(1-p)}{f_0(1-p)^2 + f_1 2p(1-p) + f_2 p^2}.$$

- Thus, p, f_0, f_1 and f_2 will enter to the likelihood calculation!

Linkage Analysis General Framework III

- Other assumptions
  - Population parameters

    e.g. $P(G)$, usually assume HWE. Need this if there are missing genotype data.

  - Single locus Inheritance probability

    e.g. $P(G_{offspring}|G_{parents,R})$, probability that a parental genotype transmits a particular allele to an offspring, usually assume Mendelian segregation.

  - Multilocus Inheritance probability if conducting multipoint linkage analysis

  - Importance of some covariates such as age and sex.

Linkage Analysis General Framework IV

- Outline of the basic analysis.
  - Parameter of interest: θ.
  - Hypotheses of interest: $H_0 : \theta = \frac{1}{2}$.
  - Other nuisance parameters related model specification: p, f_0, f_1, f_2.
  - Parameter vector: $\Phi = (p, f_0, f_1, f_2, \theta)$.
  - Data: pedigree R, phenotype Y and genotype G. (Phenotype and genotype data may be missing for some of the family members.)
  - Other data and model assumptions M: e.g. allele frequency for G and map between Gs are assumed to be known (but could be misspecified though).
  - Likelihood (could be two-point or multipoint calculations):

$$L(\Phi) = P(Y, G|R, M).$$

Linkage Analysis General Framework V

- Consider the likelihoods under H_0 and H_1.

$$L_{\tilde{\Phi}}, \tilde{\Phi} = (\tilde{p}, \tilde{f}_0, \tilde{f}_1, \tilde{f}_2, 1/2)$$

$$L_{\hat{\Phi}}, \hat{\Phi} = (\hat{p}, \hat{f}_0, \hat{f}_1, \hat{f}_2, \hat{\theta})$$

- LOD score

$$LOD = log_{10} \frac{L_{\hat{\Phi}}}{L_{\tilde{\Phi}}}$$

- For the simple Mendelian dominant or recessive disease with complete penetrance and no phenocopy, or when (p, f_0, f_1, f_2) are pre-specified based on previous studies such as population survey, segregation analysis, then $\Phi = \theta$

$$LOD = log_{10} \frac{L_{\hat{\theta}}}{L_{1/2}}$$

Limitation of Full-Parametric Linkage Analysis I

- Suitable for only simple Mendelian traits where there is a simple mode of inheritance that adequately explains the observed disease pattern in families.
- Sensitive to model misspecification. If correct: the analysis is "optimal" if the model is correct but what if the model is incorrect?
- Need large families with multiple affected individuals.

Exercises

- Chapter 6 Exercise 1.
- Chapter 6 Exercise 2.
- Chapter 6 Exercise 3.
- If we combine family 1 with grandparents unavailable with family 2 with grandmother's information,
  - Then show that then the log likelihood is

$$l(\theta) = log(\theta^1(1-\theta)^3 + \theta^3(1-\theta)^1) + log(\theta^1(1-\theta)^4).$$

  - What is the MLE?
  - What is the LOD score?
  - Show the LOD function.

What's Next

- Appendix A - Basic Concepts of (Non-Parametric) Linkage Analysis or Allele-Sharing Method
- GENEHUNTER