BioStat 710 HW4

Statistical Genetics and Genetics Epidemiology

Guangjian (Jeff) Du

Question 2 Let X_A and X_B be the number of alleles shared between two full sibs at two loci (A & B) on the same chromosome. Explain why the presence of non-zero correlation between X_A and X_B is important for nonparametric linkage analysis to work.

Answer:

Non-parametric Linkage Analysis : Based on idea that if a locus is associated with a disease, the relatives with similar phenotypes should share more alleles IBD at that locus than expected by their relatedness alone.

note #1 "non-parametric" Linkage analysis is actually full parametric assumptions. It is nonparametric in that it does not assume a model of inheritance (autosomal dominant etc)

note #2 Non-parametric linkage depends on marker loci that are close to disease locus also showing an increase in IBD relative to expectation.

This turns out to be true and it is possible to show X_A and X_B are the number of alleles shared IBD between two full sibs:

$$Cov(X_A, X_B) = \theta^2 + (1-\theta)^2 - \frac{1}{2}$$

$$Corr(X_A, X_B) = 2 * \theta^2 + 2 * (1-\theta)^2 - 1$$

Simple tests for IBD sharing in affected sib pairs compare $(\hat{P}_0, \hat{P}_1, \hat{P}_2)$ estimates of 0, 1, 2 alleles IBD to $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ these are expected IBD of 2 sibs.

Thus, the presence of non-zero correlation between X_A and X_B is important for nonparametric linkage analysis to work.

Text. section 6.4, pages 95 & 96, problem 1:
The pedigrees below show data on multiplex disease families with data at one multiallelic marker (alleles are a,b,c,d,e). Assume that the disease follows an autosomal dominant model. Assume for this part that the disease is fully penetrant $(P(disease|MMorMm) = 1)$ **and there are no phenocopies** $(P(disease|mm) = 0)$**, where M and m are the alleles at the disease locus.**

(a) Which families are phase known (the parental generation phase can be determined)?

Answer:

• If we consider the grandparents as parental generations too, then the answer would be: No family phase is known. because there's no way to determine whether the founders have MM or Mm alleles.

• If we only check parental generations, then we got scenarios below:

The phase of Family 1 is **known** :

$(G_{dad} = Ma/mc)$ and $(G_{mom} = md/md)$

The phase of Family 3 is **known** :

$(G_{dad} = Me/md)$ and $(G_{mom} = md/md)$

The phase of Family 4 is **known** :

$(G_{dad} = Mb/mc)$ and $(G_{mom} = mc/mc)$

The phase of Family 2 is **unknown** :

$(G_{dad} = Mc/md \ or \ Md/mc)$ and $(G_{mom} = md/md)$

(b) Explain why non-diseased parents are not informative for linkage.

Answer:

In order for a pedigree to contribute to the LOD score:
$LOD_{score} = log_{10}(LR(\theta))$, it must be informative – that is, each of the two loci must be heterozygous in key individuals. In an autosomal dominant disease, only the gametes of the affected parent (or non-penetrant gene carrier parent)

contribute linkage information to the LOD score. The unaffected non-gene carrier parent provides genetic information that helps interpret the transmission of the disease and marker loci to her children from the affected parent, but she does not contribute information to the LOD score.

(c) For pedigrees 1 and 3, determine the number of recombinants and nonrecombinants. Find the ML of θ based on these 2 pedigrees.

Answer:

For pedigree 1: We already know the parental phase is:

Dad (Ma/mc), Mom (md/md).

Family 1 Pedigree Table:

Parents		Dad	Ma/mc		Mom	md/md	
Offspring	Ma/md	Mc/md	mc/md	Ma/md	mc/md	mc/md	mc/md
Recombinant	s	r	s	s	s	s	s

Here, we have: r = 1, s = 6, n = 7

$\hat{\theta} = \frac{r}{n} = \frac{1}{7}$

To get the ML of θ, we need to get $L' = 0$:

$L(\theta) \propto \theta * (1-\theta)^6$

$L'(\theta) \propto [\theta * (1-\theta)^6]'$

$L'(\theta) \propto (1-\theta)^6 - 6 * (1-\theta)^5$

$L'(\theta) \propto (\theta - 1)^5 * (7\theta - 1)$

Find 2 distinct roots, only one in $0 < \theta < \frac{1}{2}$

$\theta = 1/7$

When $\theta = 1/7$, the ML = 0.056653.

For pedigree 3: We know the parental phase is:

II: Dad (Me/md), Mom (md/md).

Family 3 Pedigree Table:

Parents		Dad Me/md		Mom md/md	
Offspring	md/md	Me/md	Me/md	md/md	md/md
Recombinant	s	s	s	s	s

r = 0, s = 5, n = 5

$\hat{\theta} = \frac{r}{n} = \frac{0}{5} = 0$

$L(\theta) = \theta^0 * (1-\theta)^5 = (1-\theta)^5$

$L'(\theta) = -5 * (1-\theta)^4 = 0$

Within $0 < \theta < 0.5$, $\quad L'(\theta)! = 0$

$L(\theta)$ is a decream function within [0, 1/2], So the ML could be L(0) = 1.

When combine Family 1 and Family 3, we could have Likelihood function:

$L(\theta) = L(\theta_1) + L(\theta_3)$

$L(\theta) = \theta * (1-\theta)^6 + (1-\theta)^5$

$L'(\theta) = (\theta - 1)^4 * (7\theta^2 - 8\theta - 4)$

Get three roots, but none of them in [0, 1/2]. Only when $\theta = 0$, we could get maximum value for $L(\theta)$

So $ML = L(\theta = 0) = 1$

(d) For pedigree 4, what is the LOD score for $\theta = 0$? What would it be if the last offspring were bc rather than bb?

Answer:

For pedigree 4: One of the grandparental phase is (mm), so we know the phase is:

I: Dad (Mb/mc), Mom (md/md)

Family 4 Pedigree Table:

Parents		Dad Mb/mc		Mom md/md	
Offspring	Mb/mb	Mb/mb	mc/mb	mc/md	mb/mb
Recombinant	s	s	s	s	r

r = 1, s = 4, n = 5

$L(\theta) = \theta^1 * (1-\theta)^5$

When $\theta = 0$, $LR(\theta) = 0$;

LOD-score = $log_{10}(LR(\theta)) = log_{10}(0) = -\infty$

If the last offspring were bc rather than bb:

One of the grandparental phase is (mm), so we know the parental phase is:

Dad (Mb/mc), Mom (md/md)

Family 4 Pedigree Table under I:

Parents		Dad Mb/mc		Mom md/md	
Offspring	Mb/mb	Mb/mb	mc/mb	mc/md	mc/mb
Recombinant	s	s	s	s	s

We have: r = 0, s = 5, n = 5

$$LR(\theta) = \frac{\theta^0 * (1-\theta)^5}{(\frac{1}{2})^5}$$

When $\theta = 0$, $LR(\theta) = 2^5$;

LOD-score = $log_{10}(LR(\theta)) = log_{10}(2^5) = 1.50$

(e) Write out the likelihood for pedigree 2 as a function of θ and find the ML of θ. What is the LOD at the ML? Now assume that the phase of the affected mother is known to be Md/mc. What is the ML of and the LOD at $\theta = 0$, and LOD at $\theta = 1/2$. Comment on the effect of not knowing phase.

Answer:

We know the parental phases are ($Mm|cd$ *and* $mm|ac$).

Parental genotypes could be either of these two phases:

I: Dad (Mc/md), Mom (ma/mc)
II: Dad (Md/mc), Mom (ma/mc).

Family 2 Pedigree Table under I:

Parents		Mom Mc/md		Dad ma/mc
Offspring	ma/mc	ma/Md	mc/mc	mc/Md
Recombinant	r	r	r	r

So, under condition I, we have: r = 4, s = 0, n = 4

Family 2 Pedigree Table under II:

Parents		Mom Md/mc		Dad ma/mc
Offspring	ma/mc	ma/Md	mc/mc	mc/Md
Recombinant	s	s	s	s

r = 0, s = 4, n = 4

So the Likelihood for this pedigree is:

$$L(\theta) = \frac{1}{2} * \theta^4 * (1-\theta)^0 + \frac{1}{2} * \theta^0 * (1-\theta)^4 = \frac{1}{2} * (\theta^4 + (1-\theta)^4)$$

After derivation, we have:

$L'(\theta) \propto \theta^3 - (1-\theta)^3$

When $\theta = 0.5$, we get $L'(\theta) = 0$

When $\theta = 0.5$, we have ML of $L(\theta) = 0.125$

LOD-score $= log_{10}(LR(\theta))$

$LR(\theta = 0.5) = \dfrac{L(\theta = 0.5)}{(\frac{1}{2})^4} = 2$

LOD-score $= log_{10}(LR(\theta = 0.5)) = log_{10}2 = 0.301$

This is the LOD at the ML.

When assuming the phase of the affected Mom is Md/mc, it is under circumstance II, then the Likelihood of θ could be written:

$L(\theta) = \theta^r * (1-\theta)^{n-r} = (1-\theta)^4$

Under this condition, $L'(\theta)$ could not be zero, so there's no maximum Likelihood of θ.

When $\theta = 0$, we could have $L(\theta) = (1-0)^4 = 1$

Also, $LR(\theta) = \dfrac{(1-\theta)^4}{(1/2)^4} = 16$

LOD-score $= log_{10}LR(\theta = 0) = log_{10}16 = 1.204$

When $\theta = 1/2$, $LR(\theta) = \dfrac{(1-\theta)^4}{(1/2)^4} = 1$

LOD-score $= log_{10}(LR(\theta)) = log_{10}1 = 0$;

Comment on the not knowing phase: Here in pedigree 2, we do not know the phases of the grandparents. But we know the markers (a and d), also we know two of their offspring are unaffected, also under the full penetrance condition, so we know the Mom's genotype must be (Mm). Thus, the not knowing phases do not affect the ML and LOD calculation in this pedigree.

(f) Suppose we assume incomplete penetrance, so that $P(disease|Mm, MM) = f, 0 < f < 1$, and $P(disease|mm) = 0$. Consider pedigree 4. Can we assume the disease genotype is known for the unaffected grandmother? What can be inferred about the disease genotypes of the two parents in the middle of the pedigree? What unknown parameters, in addition to θ and f , do we need to specify in order to give an expression for the likelihood?

Answer:

Under the assumption of incomplete penetrance, we could not assume the disease genotype to be known for the unaffected grandmother.

The disease genotype of the two parents in the middle of the pedigree could be complex, the Dad could ihnerite an M allele from grandmother:

Dad (MM/bc or Mm/bc);

Also, the Mom's genotype could be:

Mom (MM/bb, Mm/bb, or mm/bb);

In order to give an expression for the likelihood, we need to know the allele frequency of M and m; Then we could get the expression of likelihood based on HWE and linkage analysis. Or, if we could get ratios of each genotype, then we could draw a hypotype table, and do the calculation based on the hypotype table.

Text. section 6.4, pages 95 & 96, problem 2:
Recall that the standard likelihood ratio test is given by

$$LRT = 2log_e LR(\hat{\theta})$$

and thus

$$maxLOD = (1/2)(log_{10}e)LRT$$

Using this connection between a standard likelihood ratio test and the maximized LOD score, show that a maximized LOD score of 3 corresponds approximately to an α-value of 0.0001.
Answer:

$$maxLOD = (1/2)(log_{10}e)LRT = 3LRT = \frac{3}{(1/2)(log_{10}e)} = 13.8155$$

Since LRT follows χ^2 distribution with degree of freedom 1.

For LRT = 13.855, we have p-value = 0.0002.

So it corresponds to an approximate α-level of 0.0001.

Text. section 6.4, pages 95 & 96, problem 3:
Show that for a given value of θ, we can combine information for independent pedigrees by simply summing the LOD scores evaluated at θ. Hint, if the pedigrees are independent, then the likelihood of all the pedigrees is obtained by multiplying the likelihood of each separate pedigree.

Answer:

Since the pedigrees are independent, we could assume for n pedigrees, each likelihood would be written:

$LR(\theta_i), i = 1, 2, 3, , , n.$

By multiplying the likelihood of each separate pedigree, we could get the likelihood of all pedigrees:

$$LR_{total} = \prod_{i=1}^{n} LR(\theta_i)$$

Since we have LOD-score = $log_{10}(LR(\theta))$

The total LOD-score could be written in:

LOD-score-total = $log_{10}(LR_{total})$

$$= log_{10}[\prod_{i=1}^{n} LR(\theta_i)]$$

$$= \sum_{i=1}^{n} log_{10}[LR(\theta_i)]$$

$$= \sum_{i=1}^{n} LOD(\theta_i)$$

So, the combined LOD score for independ pedigrees could get by summing the LOD scores of each pedigree evaluated at θ.