ゲノム統計学
―連鎖解析と関連分析―

東京大学大学院医学系研究科
クリニカルバイオインフォマティクス研究ユニット
田中　紀子

講義内容

- 連鎖解析
  - パラメトリックな方法（ロッド値法）
  - ノンパラメトリックな方法（Affected sib pair method:ASP)
- ケース・コントロール関連分析
- 伝達不平衡試験
（transmission disequilibrium test:TDT)

連鎖分析

- ある疾患の原因遺伝子が染色体上のどの位置に存在するかを解析する手法。

- ある疾患の原因遺伝子の近傍に位置するDマーカーは家系内では疾患とともに遺伝する（連鎖している）ことを利用する。

ロッド値法（Morton,1955)

- ロッドスコア関数

$$Z(\theta) = \log_{10}[L(\theta)/L(1/2)]$$

L(θ)＝ある遺伝子型の下である表現型が観測される尤度　θ：組換え率

から計算される値をロッドスコアという。

ロッドスコア＞３で自由組換えを棄却（連鎖している）、ロッドスコア＜－２で自由組換えを採択（連鎖していない）。

つまり、

帰無仮説：マーカー（遺伝子）と疾患感受性遺伝子は連鎖していない $(\theta = 0.5)$

という仮説に対して

対立仮説：マーカー（遺伝子）と疾患感受性遺伝子は連鎖している $(\theta \neq 0.5)$

という仮説を尤度比検定しているのに等しい

ロッドとは？

- Logarithm of Oddsの略
- オッズ
  - オッズ・・・確率の比

$$\frac{p}{1-p}$$

  - $$\frac{\text{（対立仮説（連鎖している）下での尤度）}}{\text{（帰無仮説（連鎖していない）の下での尤度）}}$$

- ロッドスコアの場合習慣的に常用対数をとっている

尤度

- 尤度関数:
  - 観測データが確率密度関数f(x; θ)となるある分布から独立に抽出されたという仮説の下で、θ の各値が母数の真の値であることの尤もらしさを示す関数

$$\mathrm{L}(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

連鎖解析の尤度

- 観測されたn人の個体の表現型$(x_1,\cdots,x_n)$の現れる確率を$P(x_1,\cdots,x_n)$、観測された個体の遺伝子型の確率$P(g_1,\cdots,g_n)$とすると、ある遺伝子型のもとで観測された表現型の得られる確率(penetrance:浸透率）は独立であることから、この家系データの尤度は 一般に

$$L = P(x,g) = \sum_g p(x|g)p(g)$$

ただし $\quad p(x|g) = \prod_i^n p(x_i|g_i), \quad p(g) = P(g_1,\cdots,g_n), \quad i = 1,\ldots,n$

となり、遺伝子型は組換え率θの関数として表わすことが出来るので、

$$L(\theta) = \sum_g P(g|\theta)P(x|g)$$

で計算される

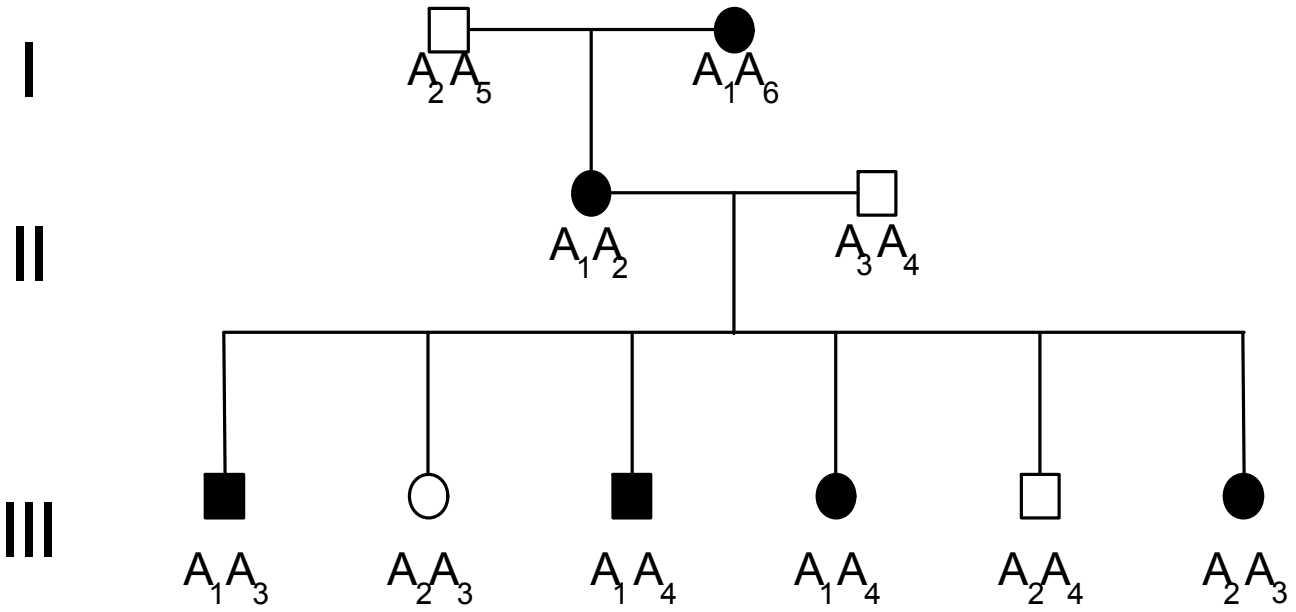
最尤推定量

- 母数 θ の真値
  - 尤度関数　$L(\theta)$が最大になる θ の値

- 最尤推定量
  - Maximum Likelihood Estimator (MLE)
  - 多くの場合　$\hat{\theta}(X)$

例1-すべての相が分かっている場合-

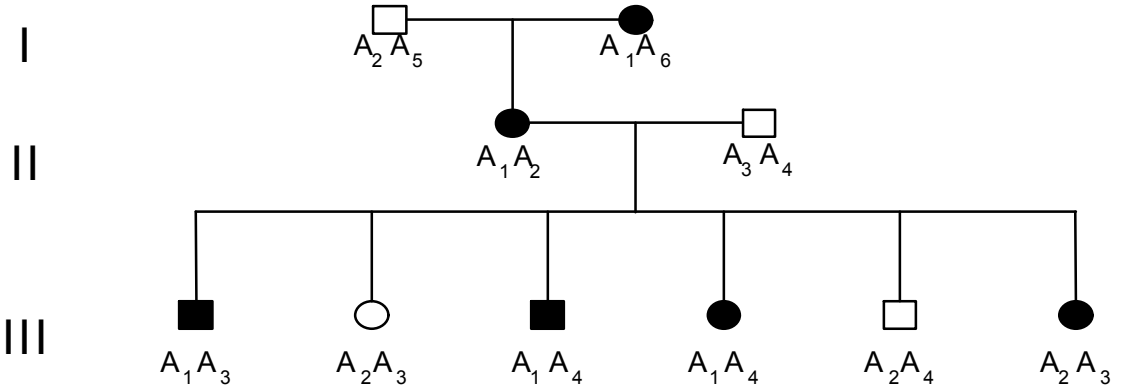
- 次のような家系のデータが得られたとする



(Strachan & Read,1999)

仮に、対立遺伝子A1と疾患が連鎖していて、疾患が優性形質であるとすると、遺伝子座Aと疾患感受性遺伝子は連鎖しているでしょうか？

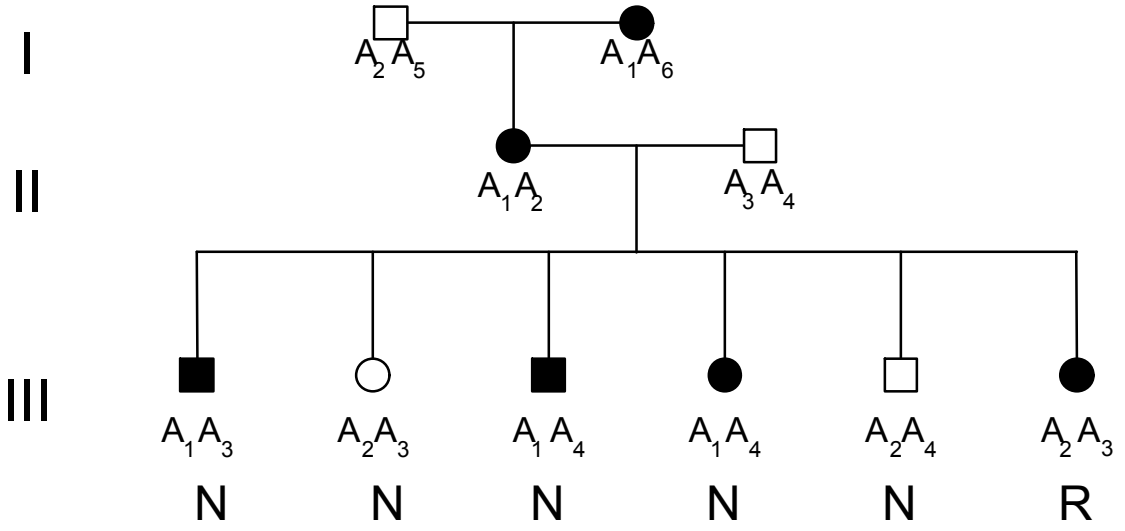
例1（続き）



- いま、A1の近傍に疾患感受性遺伝子があると仮定すると…
  - 組換えが起きなければ、II1からA_1を受け継いだ子供は全て患者となり、A_2を受け継いだ子供は患者にはならないはず
  - 組換えが起きればA_2を受け継いだ子供は患者になり、A_1を受け継いだ子供は患者にならないはず

と、なるので

例1（続き）



尤度関数：

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

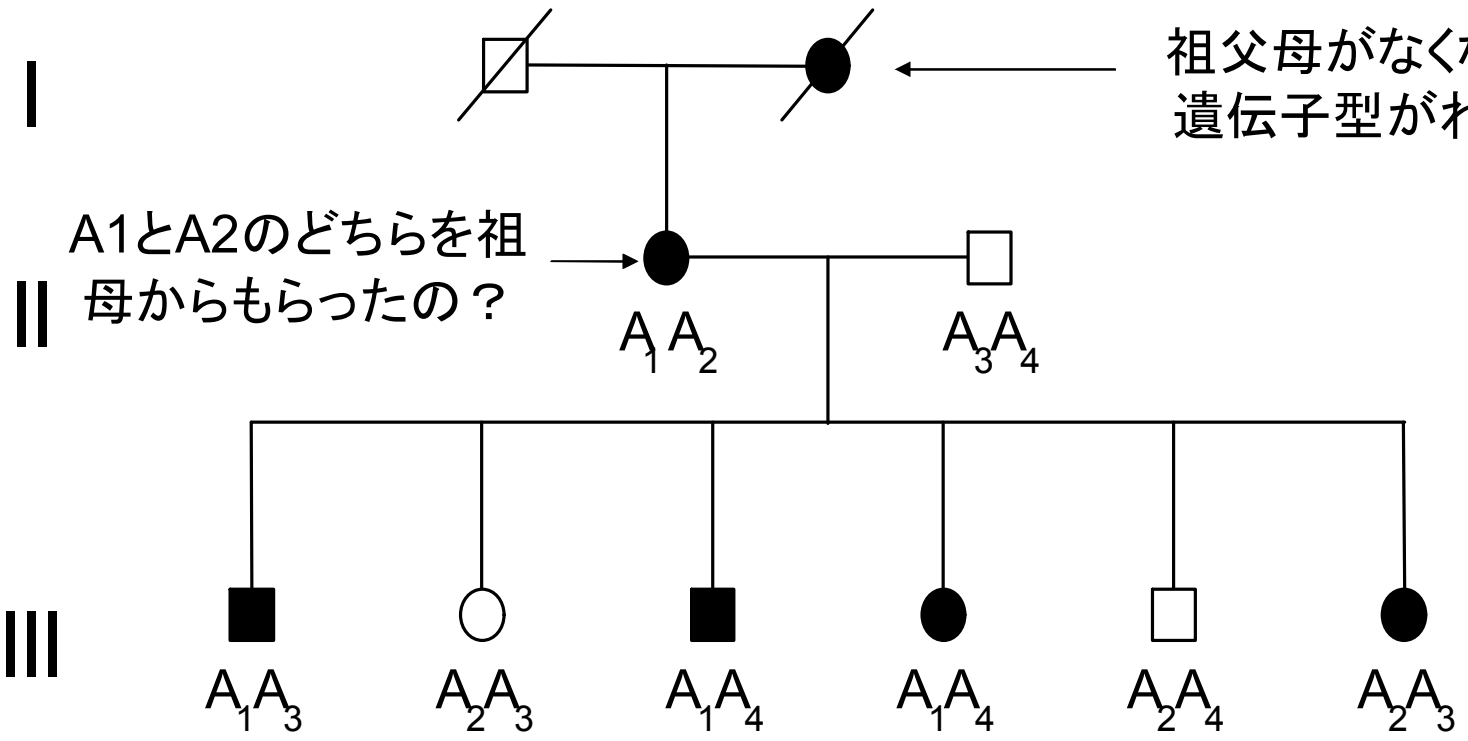
ロッドスコア関数：

$$Lod(\theta) = \log_{10}\left[(1-\theta)^5 \theta / (1/2)^6\right]$$

このあたりで尤度関数が最大に（正確には1/6)

θ	0	0.1	0.2	0.3	0.4	0.5
LOD(θ)	$-\infty$	0.577	0.623	0.509	0.299	0

例2-親の相がわからないとき-

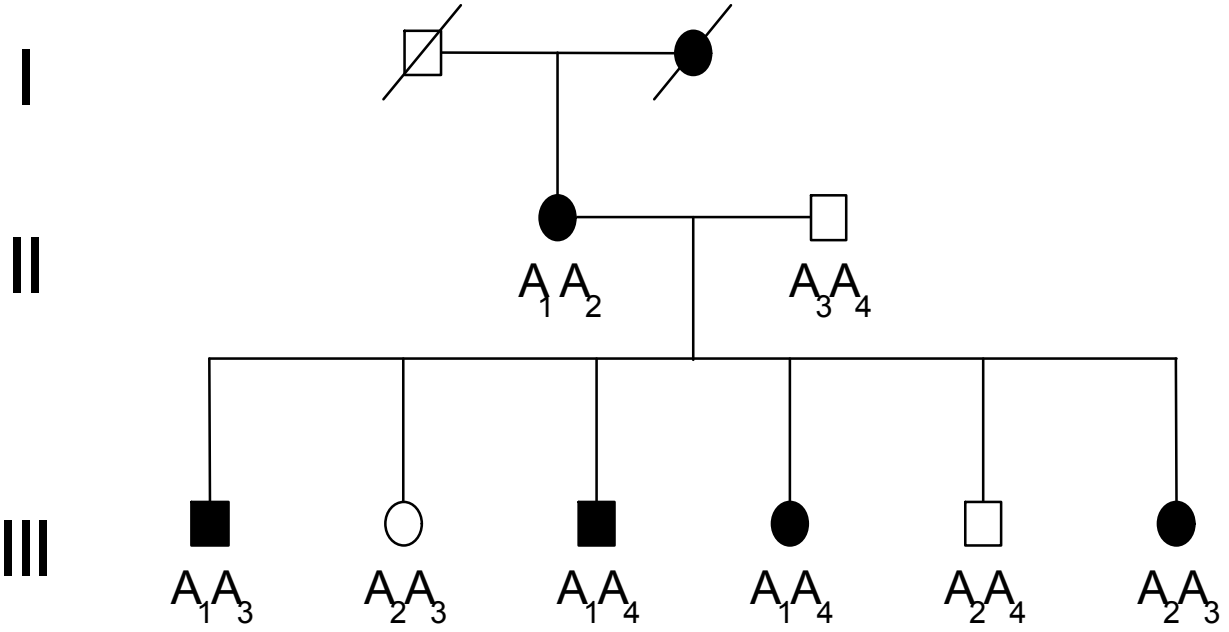


(Strachan & Read 1999)

子供が組換え型か非組換え型かわからない！

対立遺伝子A1かA2は疾患感受性遺伝子と連鎖しているでしょうか？

例2-続き-



A1が疾患感受性遺伝子と連鎖している場合 → N N N N N R

A2が疾患感受性遺伝子と連鎖している場合 → R R R R R N

尤度関数: $\mathrm{L}(\theta)=\frac{1}{2}\times(1-\theta)^5\theta+\frac{1}{2}\times\theta^5(1-\theta)$

ロッドスコア関数: $\mathrm{Lod}(\theta)=\log_{10}\left[\frac{1}{2}\times(1-\theta)^5\theta\Big/(1/2)^6+\frac{1}{2}\times(1-\theta)\theta^5\Big/(1/2)^6\right]$

例1と例2の計算結果の比較

例2

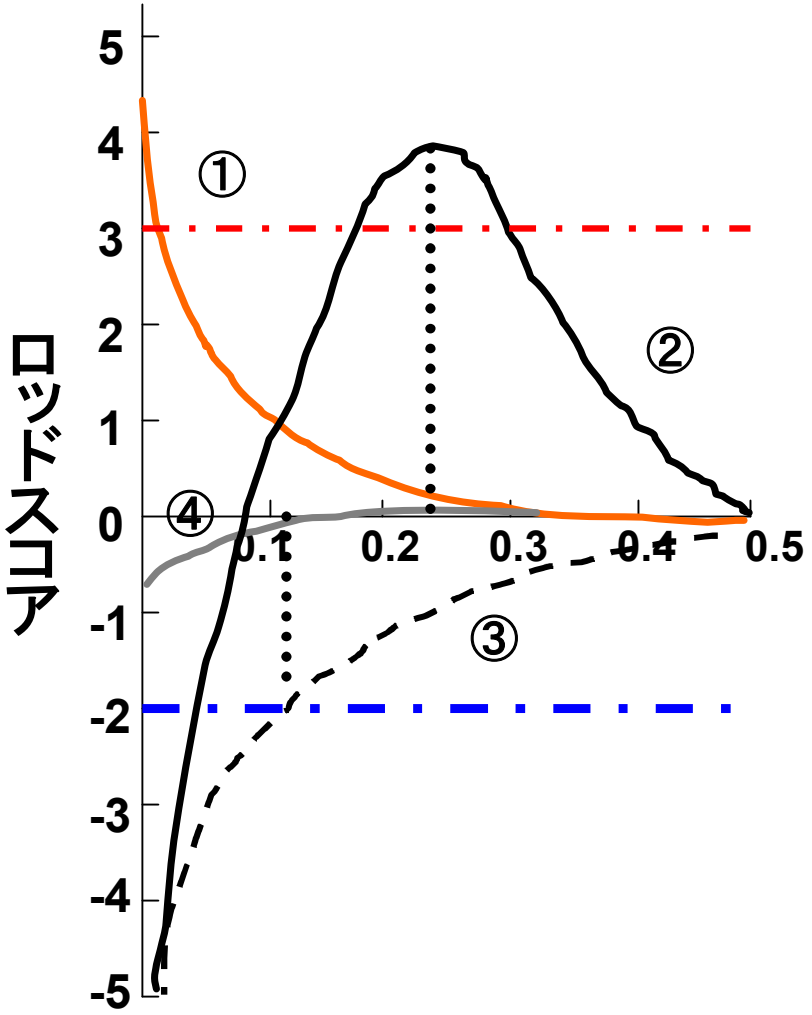
θ	0	0.1	0.2	0.3	0.4	0.5
LOD(θ)	$-\infty$	0.276	0.323	0.222	0.076	0

例1

θ	0	0.1	0.2	0.3	0.4	0.5
LOD(θ)	$-\infty$	0.577	0.623	0.509	0.299	0

情報が減った（親の相がわからなかった）ことにより解析感度が下がったことがわかる

ロッドスコア曲線



- 曲線①

組換え型がいない家系

- 曲線②

組換え率＝0.23で推定される家系

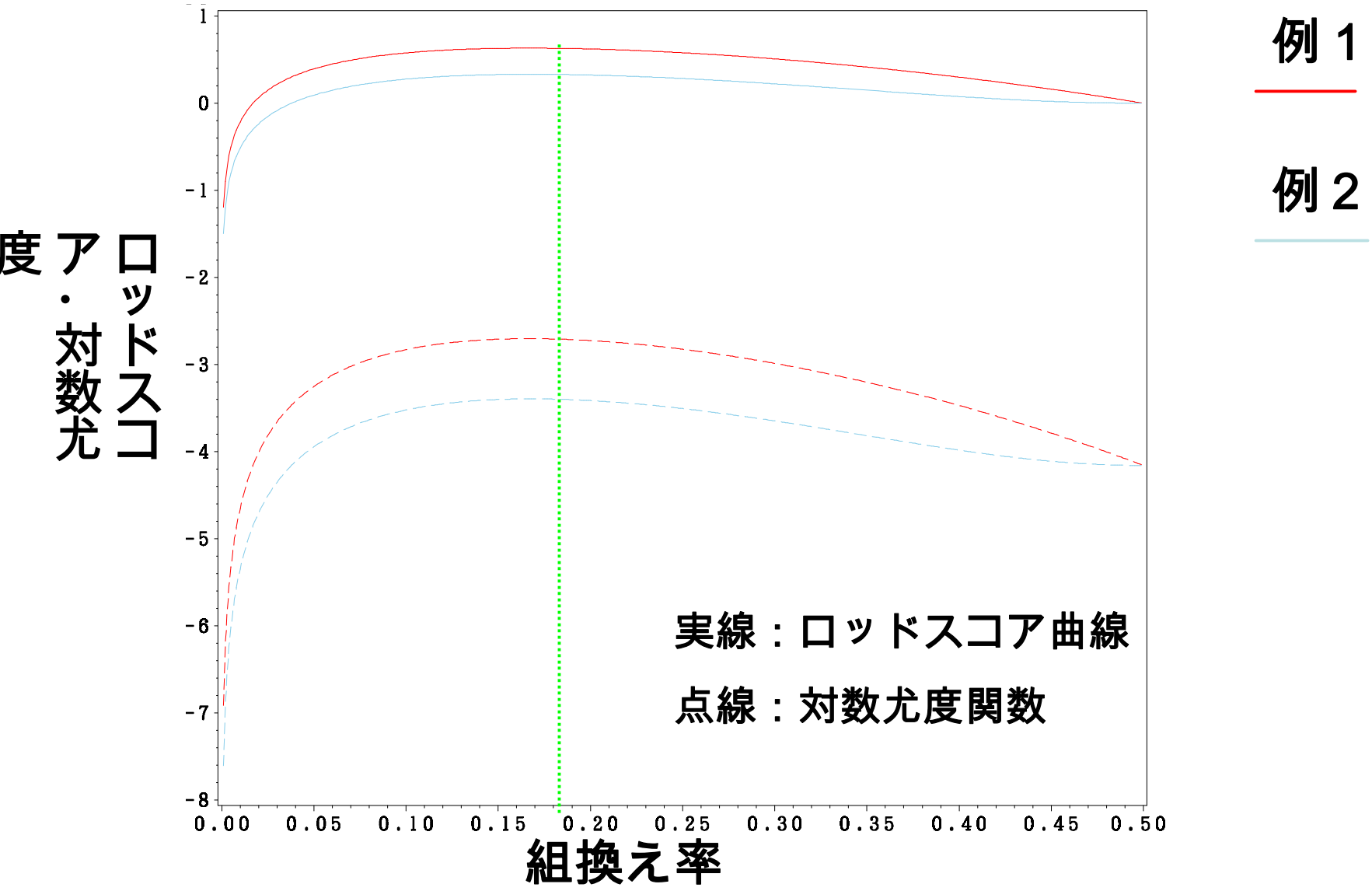
- 曲線③

組換え率0.12以下で“連鎖していない”となる家系

- 曲線④

なんとも結論できない家系

例題のロッドスコア曲線と対数尤度関数



ロッドスコアの意味

- Lodスコアが3.0以上
  ＝偶然に対応関係がみられるよりも1000倍以上　確率の高い場合に検定で有意差ありとする。
  ＝尤度比検定した場合の有意水準を0.0001にする。

- 3.0より小さい場合には否定的というのではなく、1.0～2.0はinteresting、2.0～3.0はsuggestiveとする場合もある。

罹患同胞対法
（Affected sib pair method:ASP　method)

- 同じ疾患に罹患した兄弟で観察された共有する同祖遺伝子（alleles identical by descent: IBD）の割合が、連鎖がないと仮定した場合に期待される割合から有意に偏っているかどうかを検定する方法
- 遺伝様式を仮定しなくても検定することができるので、ノンパラメトリックな方法と呼ばれる
- そこで、ロッド値法などパラメトリックな方法に比べ、多因子疾患やありふれた疾患の研究に適用しやすい

ASP法の検定方法

- 帰無仮説：疾患感受性遺伝子と遺伝子座Aにある対立遺伝子が連鎖していない

（罹患同胞対のあいだで共有するIBDの平均期待割合＝0.5）

- 対立仮説：疾患感受性遺伝子と遺伝子座Aにある対立遺伝子が連鎖している

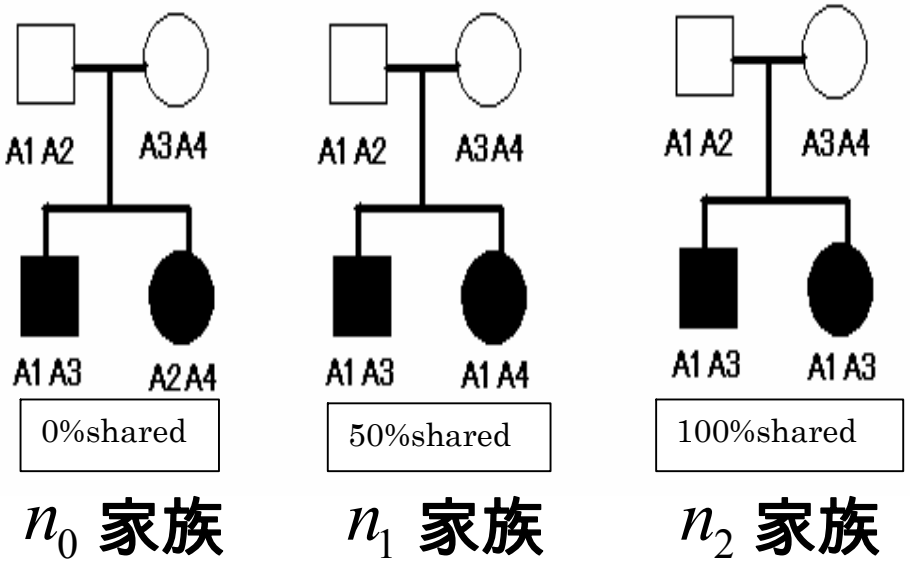
（罹患同胞対のあいだで共有するIBDの平均期待割合＞0.5）

として、平均値の差の検定を行う

ASP法の検定統計量

- 共有するIBDの数をX、X=0,1,2でそれぞれの観察出現家系頻度を
  $n_0, n_1, n_2 (n = n_0 + n_1 + n_2)$

とすると、帰無仮説（連鎖していない）の下での期待IBD共有割合は0.5なので、検定統計量は



n_0 家族　n_1 家族　n_2 家族

$$T_{ASP} = \frac{\left(\frac{n_1 + 2n_2}{2n} - \frac{1}{2}\right)}{s\sqrt{n}},$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=0,1,2}\left(\frac{n_1 + 2n_2}{2n} - e_i\right)^2 n_i},$$

$$(e_0, e_1, e_2) = (0, 0.5, 1)$$

と計算され、これは自由度(n-1)のt分布に従うことからp値を計算する

例
-インシュリン依存型糖尿病（IDDM）とIDDM4-

- IDDM
  - Ｉ型糖尿病（Type 1 diabetes)ともいわれ、**膵臓のランゲルハンス島にあるβ細胞が破壊されてインシュリンを分泌する機能そのものがなくなってしまうタイプの糖尿病**
  - 多くの場合、若年発症で、インシュリン療法が主に行われる
  - NIDDM(インシュリン非依存型糖尿病）と比べてより、家族集積性が強い

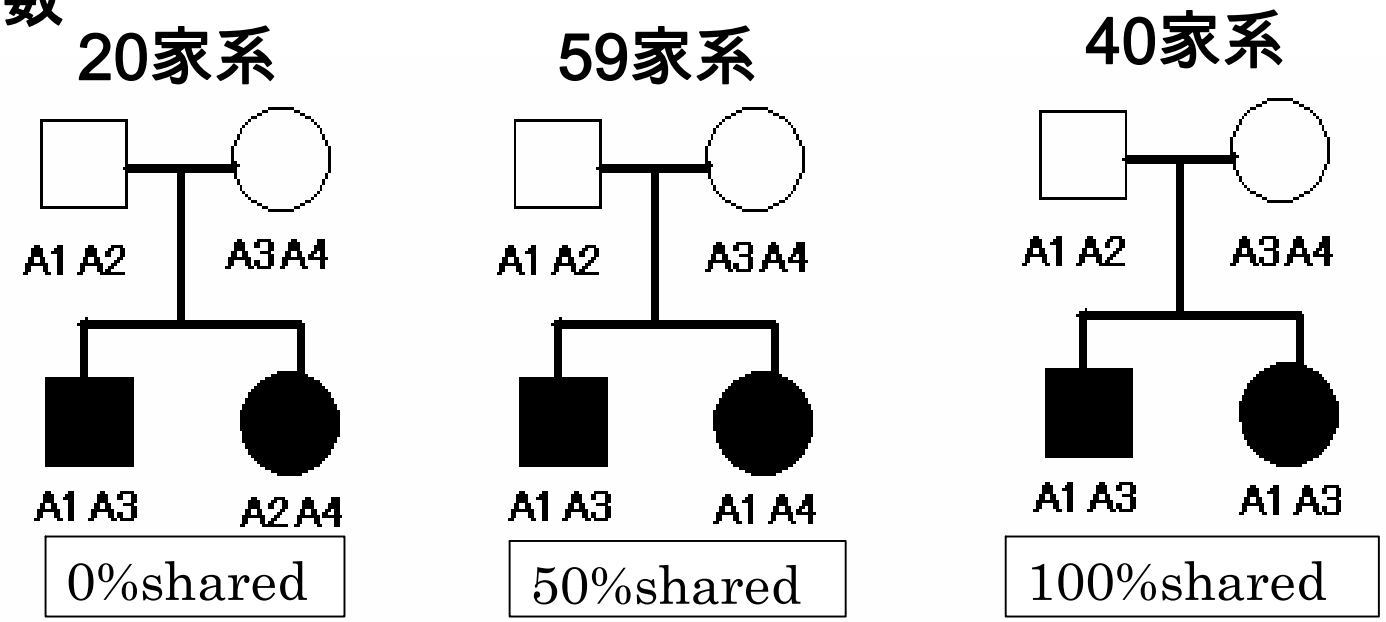
例の続き
-インシュリン依存型糖尿病（IDDM）とIDDM4-

- Hashimotoら（1994）の研究（IDDM4のmapping)
  - IDDM患者家族を対象に染色体11q13のFGF 3(Fibroblast growth factors 3:線維芽細胞増殖因子) 座位での罹患同胞対調査を行ったところ、119同胞対中、共有するIBDが0,1,2個の同胞対数はそれぞれ、20,59,40,であった。

FGF3座位とIDDM感受性遺伝子は連鎖しているでしょうか？

計算例

観察数



平均同祖遺伝子共有の割合

$$\frac{0\times20+0.5\times59+1.00\times40}{(20+59+40} = 0.58\ (\ SD=0.346\ )$$

t値 = (0.58-0.5) / (0.346/$\sqrt{119}$)

= 2.52

→ 対応する p =0.0058

(v = 118)

ケース・コントロール関連分析

✧ある疾病の患者（ケース）群と対照（コントロール：その疾病に罹患していない者）群を設定し、過去にさかのぼって仮説的要因の曝露率などを比較して要因と疾病の関連性を調べる方法

- 連鎖解析よりせまい領域に疾患感受性遺伝子座を絞り込むことができる
- 多因子疾患・ありふれた疾患に有効な方法

-GRR(genotype relative risk)-

２つの対立遺伝子A,aのある遺伝子座について、ある集団である疾患の浸透率を調べると、得られるデータは下のように要約できる。

	Genotype		
	AA	Aa	aa
ある疾患に罹患している	p_{AA}	p_{Aa}	p_{aa}
ある疾患に罹患していない	$1-p_{AA}$	$1-p_{Aa}$	$1-p_{aa}$

GRRは $\theta_{AA}=\frac{p_{AA}}{p_{aa}}$ $\theta_{Aa}=\frac{p_{Aa}}{p_{aa}}$

この疾患には、遺伝子型AAの人はaaの人に比べて θ_{AA} 倍リスクが高い（低い）

この疾患には、遺伝子型AAの人はaaの人に比べて θ_{Aa} 倍リスクが高い（低

関連の指標

- オッズ比
  - オッズ・・・確率の比 $\dfrac{p}{1-p}$
  - $\dfrac{\text{（ある疾患にかかる確率）}}{\text{（ある疾患にかからない確率）}}$

▶オッズ比・・・オッズの比 $\dfrac{\dfrac{p}{1-p}}{\dfrac{q}{1-q}}$

ある遺伝子型に対するほかの遺伝子型の、相対的な病気にかかりやすさを示す指標（つまり相対的な関連の強さの指標）となる。

関連の指標-GRR(genotype relative risk)-

２つの対立遺伝子A,aのある遺伝子座について、ある疾患のケース･コントロール研究を行うと、得られるデータは下のように要約できる。

	Genotype			
	AA	Aa	aa	total
Case	γ_{AA}^{1}	γ_{Aa}^{1}	γ_{aa}^{1}	R
control	γ_{AA}^{0}	γ_{Aa}^{0}	γ_{aa}^{0}	S

人

オッズ比は

$$\theta_{AA}^{*} = \frac{\gamma_{AA}^{1}}{\gamma_{aa}^{1}} \Big/ \frac{\gamma_{AA}^{0}}{\gamma_{aa}^{0}} \qquad \theta_{Aa}^{*} = \frac{\gamma_{Aa}^{1}}{\gamma_{aa}^{1}} \Big/ \frac{\gamma_{Aa}^{0}}{\gamma_{aa}^{0}}$$

$\gamma_i^0 \approx (\gamma_i^0 + \gamma_i^1)$ の場合

$$\theta_{AA} \approx \theta_{AA}^{*}, \quad \theta_{Aa} \approx \theta_{Aa}^{*}$$

-HRR(Haplotype relative risk)-

- Multiplicative model（つまり、$\theta_{AA}=\theta_{Aa}^2$ ）の下では、特別に$\theta_{Aa}=\psi$ をHRR(haplotype relative risk)といい、H-W平衡の下で対立遺伝子頻度と次のような関係が導き出せる。

$$\omega_A^{case}=\frac{\psi\omega_A^{pop}}{\omega_A^{pop}+\psi\omega_A^{pop}},\quad \omega_a^{case}=\frac{\omega_a^{pop}}{\omega_a^{pop}+\psi\omega_A^{pop}}$$

ω_i^{case} CaseのAllel iの頻度

ω_i^{pop} 集団のAllel iの頻度

$\omega_i^{control}$ ControlのAllel iの頻度

$\omega_i^{pop}\approx\omega_i^{control}$ の場合

$$\psi\approx\psi^*=\frac{\omega_A^{case}}{\omega_a^{case}}\Big/\frac{\omega_A^{control}}{\omega_a^{case}}$$

関連の指標-HRR(Haplotype relative risk)-

２つの対立遺伝子A,aのある遺伝子座についてケース・コントロール研究した結果、得られるデータは下のようにも要約できる。

	Allele		
	A	a	total
case	r_A	r_a	R
control	s_A	s_a	S

(chromosomes)

ある疾患に関して、対立遺伝子Aのaに対する発症リスクは

$$\psi^* = \frac{\frac{r_A}{R}}{\frac{r_a}{R}} \Big/ \frac{\frac{s_A}{S}}{\frac{s_a}{S}} = \frac{r_A s_a}{s_A r_a}$$

倍と計算される

オッズ比の分散

- 一般に、下のような表でデータが要約されたとき、そこから計算されるオッズ比と対数オッズ比の漸近的分散は

$$OR = \frac{ad}{bc} \qquad Var(\ln(OR)) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

となるので、オッズ比の95%信頼限界は

$$\exp\left(\ln(OR) \pm 1.96 \times \sqrt{Var(\ln(OR))}\right)$$ **と計算される**

	Exposure	
	Yes	No
case	a	b
control	c	d

例-乳がんとBRCA1-

- 乳がん
  - がんの中でもcommon（ヨーロッパ・アメリカにおいて成人女性で生涯リスク10%前後）
  - 40—60歳代で発症
  - 死亡リスクは他のがんに比べて低い
- BRCA1
  - 1990年に17q21にマップされる
  - 変異があると70歳くらいまでに80—90%くらい乳がんに発症すると報告されている

例の続き-BRCA1と乳がん-

- Danningらの研究（1997）
  - BRCA1遺伝子にあるアミノ酸塩基置換を起こす変異の中でも多型頻度の比較的高いPro871Leuについて、乳がんとの関連を調べるためのpopulation based case-control study
  - ケース800人、コントロール572人について、タイピング。

1. LeuはProより乳がん発症のリスクが高いか？

2. Leu/Leu , Pro/LeuはPro/Proより乳がん発症リスクが高いか？

例の続き-BRCA1と乳がん-

	Allele		
	Leu	Pro	total
case	*547*	*1053*	*1600*
control	*362*	*782*	*1144*

Haplotype relative risk は

$$\hat{\psi}^* = \frac{547}{1053} \Big/ \frac{362}{782} = 1.122$$

95%CI:(0.95, 1.32)

Pro/Pro遺伝子型をreferenceとするとGenotype relative riskは

	Genotype			
	Leu/Leu	Leu/Pro	Pro/Pro	total
case	*89*	*369*	*342*	*800*
control	*56*	*250*	*266*	*572*

$$\hat{\theta}^*_{LL} = \frac{89}{342} \Big/ \frac{56}{266} = 1.236$$

95%CI:(0.85, 1.79)

$$\hat{\theta}^*_{LP} = \frac{369}{342} \Big/ \frac{250}{266} = 1.148$$

95%CI:(0.92, 1.44)

解析方法—検定方法1-

対立遺伝子頻度としてデータを要約した場合

	Allele		
	A	a	total
case	r_A	r_a	R
control	s_A	s_a	S
total	n_A	n_a	N

(chromosomes)

対象とした疾患と対立遺伝子A,aが

関連があるかどうか知りたい

→ **仮説検定**

検定方法1の続き

	Allele		
	A	a	total
case	$r_A(p_{11})$	$r_a(p_{12})$	R
control	$s_A(p_{21})$	$s_a(p_{22})$	S
total	n_A	n_a	N

帰無仮説： $p_{ij} = p_{0ij}$

対立遺伝子A, aと疾患とに関連はない

対立仮説： $p_{ij} \neq p_{0ij}$

対立遺伝子A, aと疾患とに関連がある

この場合、検定は割合の差の検定、もしくはpearson χ 2乗検定を行い、検定統計量は

$$\chi^2{}_p = \frac{N(r_A s_a - r_a s_A)^2}{RSn_A n_a}$$

自由度１のχ2乗分布に従うとしてp値を計算する

解析方法-検定方法2-

遺伝子型頻度として、特に対立遺伝子Aに興味があってデータを要約した場合

	A allele			
	0	1	2	total
case	r0	r1	r2	R
control	s0	s1	s2	S
total	n0	n1	n2	N

対立遺伝子Aを多く持っているほど対象とした疾患と関連があるかどうか知りたい

検定方法2の続き

帰無仮説：対立遺伝子A, aと疾患とに関連はない

対立仮説：対立遺伝子Aと疾患とに線形の関連がある

この場合の検定は対立仮説が線形性の検出に絞られているので、Armitageの傾向検定を行い、検定統計量は

$$Y^2 = \frac{N\{N(r_1 + 2r_2) - R(n_1 + 2n_2)\}^2}{R(N - R)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}}$$

自由度１のχ2乗分布に従うとしてp値を計算する

例-BRCA1と乳がん-

	Genotype			
	Leu/Leu	Leu/Pro	Pro/Pro	total
case	*89*	*369*	*342*	*800*
control	*56*	*250*	*266*	*572*

Pro/Pro遺伝子型をreferenceとするとGenotype relative riskは

$$\hat{\theta}^{*}_{LL} = \frac{89}{342} \Big/ \frac{56}{266} = 1.236$$

$$\hat{\theta}^{*}_{LP} = \frac{369}{342} \Big/ \frac{250}{266} = 1.148$$

95%CI:(0.85, 1.79)

95%CI:(0.92, 1.44)

傾向性の検定結果は

$Y^2 = 1.98$, **P = 0.16**

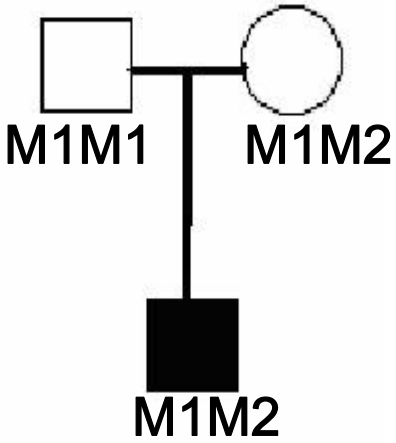
つまり、Pro871Leu多型とIDDMとに関連はないことは否定できない

伝達不平衡試験
（transmission disequilibrium test :TDT)

- Spielman（1993）らは関連（例えば連鎖不平衡）の存在下で連鎖の有無を検定する方法としてTDTを提案
- 連鎖の有無だけではなく、population stratification の存在があっても連鎖不平衡による関連を検出できる
- 大きな家系は必要なく、病気の子供一人とその両親の遺伝情報を必要とする。

TDTの検定統計量

n家系サンプルした場合



	遺伝しなかった対立遺伝子		
遺伝した対立遺伝子	M1	M2	総数
M1	a	b	$a+b$
M2	c	d	$c+d$
総数	$a+c$	$b+d$	$2n$

もし、M1が病気と関連があれば（感受性遺伝子であれば）、M1を伝えた場合の方がM2を伝えた場合よりも多いはず

帰無仮説 $H_0: p_{1.} = p_{.1}$ (M1を伝える確率とM2を伝える確率は等しい）を検定する

TDTの検定統計量はマクネマー検定統計量に一致し、

$$\chi^2_{TDT} = (b-c)^2 / (b+c)$$

例

100人のある疾患Dに罹患した子供とその両親を対象に、二つの対立遺伝子A, aからなる候補遺伝座Aと疾患Dとの関連を調べる研究をおこなった。

すると、子供と両親の遺伝子型は次のようになった。

	子供		
両親	A/A	A/a	a/a
AA × AA	*22*	*0*	*0*
AA × Aa	*17*	*25*	*0*
AA × aa	*0*	*7*	*0*
Aa × Aa	*1*	*11*	*13*
Aa × aa	*0*	*1*	*1*
aa × aa	*0*	*0*	*2*

(Sham P.1999)

例—続き—

両親	子供 A/A	A/a	a/a
AA × AA	22	0	0
AA × Aa	17	25	0
AA × aa	0	7	0
Aa × Aa	1	11	13
Aa × aa	0	1	1
aa × aa	0	0	2

H0:二つの対立遺伝子の伝達確率は同じ

を検定する。検定統計量は

$$\chi^2_{TDT} = (31-63)^2/(31+63)$$
$$= 10.89$$

これが自由度１のχ2乗分布に従うので、p = 0.00097

つまり、0.1%の水準でも帰無仮説は棄却され、伝達確率が異なる可能性が示された。

このデータは次のようにも要約することができる

伝わった	伝わらなかった A	a	total
A	22 ×2+17+25+7=93	17+1 ×2+11+1=31	124
a	25+11+13 ×2+1=63	7+1+1+2 ×2=13	76
total	156	44	200

まとめ

- 連鎖解析
  - パラメトリックな方法（ロッド値法）
  - ノンパラメトリックな方法（Affected sib pair method:ASP)
- ケース･コントロール関連分析
- 伝達不平衡試験

（transmission disequilibrium test:TDT)

状況に応じて、これらのデザイン・解析手法を選択することが必要

参考図書

- Rice JA. (1994)" Mathematical Statistics and Data Analysis " ,Thomson Learning
- Rothman KJ, Greenland S.(1998)"Modern Epidemiology-2nd ed.", LW&W,Philadelphia
- Armitage P, Berry G.(1994)"Statistical methods in Medical Research-3rd ed.",Blackwell Science, Baltimore and London
- Strachan T, and Read A. (1999)"Human Molecular genetics 2", BIOS
- Sham P.(1998)" Statistics in Human Genetics", John Wiley & Sons, NY
- Balding DJ, et al.(2001)"handbook of Statistical Genetics", John Wiley & Sons, NY

参考文献

- Hashimoto L, et al.1994.Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature 371: 161-4.
- Dunning. A.,et al.1997.Common BRCA1 variants and susceptibility to breast and ovarian cancer in the general population. Hum. Mole. Genet. 6:285-9.
- Spielman RS, et al.1993.Transmission test for linkage disequilibrium: The Insulin gene region and Insulin-dependent diabetes mellitus(IDDM). Am. J. Hum. Genet. 52:506-16.