遺伝統計学入門[(1)]・補遺集

株式会社スタージェン

(1) 本冊子は東京女子医科大学附属膠原病リウマチ痛風センター所長・鎌谷直之教授著・遺伝統計学入門（2007，岩波書店）の補遺集です

Penhaplo と QTLhaplo

第 7 章 ― 7.6 ハプロタイプ推定のアルゴリズム ―

1 Penhaplo と QTLhaplo（第 7 章、7.6）

1.1 ハプロタイプと SNP を統一的に取り扱う標本空間

我々は以下のように、ハプロタイプと単座位（例えば SNP）のアレルを統一的に取り扱うため次の標本空間と出来事（event）を定義する。そして、その標本空間上の出来事と表現型との関連を解析するアルゴリズムを構築する。

ここで、完全ハプロタイプとは、特定の染色体領域のすべての多型座位のアレルのリスト（その領域のハプロイドゲノムのすべての情報)、不完全ハプロタイプとは、その限られた座位のアレルのリストである。

Ω: 全ての完全ハプロタイプの集合

H_i : i 番目の完全ハプロタイプ

H_i : (再定義): その要素が H_i である Ω の部分集合.

X : 一つの SNP 座位の一つのアレル

X (再定義): X をリストの要素として持つすべての完全ハプロタイプの集合

A_i: 限られた座位、例えば htSNP 座位のみに特定のアレルを持つ i 番目のハプロタイプ

A_i (再定義): 特定の座位（複数可）（例えば htSNP 座位）が A_i である完全ハプロタイプの集合（不完全ハプロタイプ）

ここで、

$\{X\} \subset \{A_i\} \subset \{\{H_i\}\} = \{0,1\}^{\Omega}$

であることがわかる。

即ち、すべての不完全ハプロタイプの集合はすべての完全ハプロタイプの集合の集合（即ち Ω のべき集合）と同じではなく、それより小さい。

こうして、完全ハプロタイプ、不完全ハプロタイプ、単座位のアレルを同じ標本空間上の出来事として定義しできた。個人に関するハプロタイプの情報は二つのハプロタイプの組み合わせであるディプロタイプ形であるが、ディプロタイプ形は二つの完全ハプロタイプの組み合わせである。上の定義で不完全ハプロタイプ A_i、単座位の SNP のアレル X を Ω の部分集合として定義した。従って、特

定のディプロタイプ形は特定の不完全ハプロタイプ A_i、または単座位の SNP のアレル X の要素を何個保有しているかにより 0,1,2 個に相当する状態（出来事）に分類される。

ディプロタイプ形により質的表現型を発現する確率（浸透率）が決まり、量的表現型の分布（分布のパラメータ）が決まるというのが遺伝継承法則（優劣の法則）をハプロタイプに拡張して得られる法則である。

以上の考えに基づき、ディプロタイプ形と質的表現型の関連を検定し、浸透率を推定するアルゴリズムが PENHAPLO であり、ディプロタイプ形と量的表現型の関連を検定し、分布のパラメータを推定するアルゴリズムが QTLHAPLO である。

これまでのハプロタイプと表現型の関連の検定では、case と control の特定のハプロタイプの頻度（frequency）が比較される事が多かった。もちろん、そのような頻度の比較も大切であるが、これは必ずしも遺伝継承法則に基づいたものではない。PENHAPLO と QTLHAPLO はハプロタイプではなく、個体のディプロタイプ形に基づいた、遺伝情報と表現型の検定ができる点が重要である。また個体レベルのアルゴリズムであるため浸透率や分布のパラメータが推定できる点も優れている。

遺伝学的問題の場合、推定や検定に用いられる手法はほとんどの場合最尤法である。これは遺伝継承法則が確率関数として定義されている事による。遺伝学的解析に尤度や最尤法が多く用いられるのは当然のことである。尤度の概念を導入したのは Fisher であるが、Fisher は遺伝的問題の解決のため（具体的には連鎖解析）遺伝継承に基づいた最尤法を導入したのである。

1.2 ディプロタイプ形と質的表現型の関連を検定し、浸透率を推定する PENHAPLO

ハプロタイプ推定をオーダーメイド医療に結びつけるためには遺伝的情報と表現型情報の関係を検定する手法が必要である。我々は遺伝子型データと表現型データを観察データとし、集団のハプロタイプ頻度とハプロタイプを基礎とした浸透率をパラメータとし、ハプロタイプの有無と表現型の関係を検定し、浸透率を推定するアルゴリズム PENHAPLO を構築した [1] [2]。

1.2.1 一般的なハプロタイプ推定の標本空間

遺伝子型データを用いた一般的なハプロタイプ推定のアルゴリズムにおいては、標本空間は次のような試行（あるいは実験）の結果の集合であると定義される。最初に無限のハプロタイプコピーの集合におけるハプロタイプ頻度を与える。そのハプロタイプ頻度に基づいて N 人の個体のそれぞれにハプロタイプコピーの集合から 2 つのハプロタイプを順番に引き出し与える。観察データはすべての個体のハプロタイプに関係するすべての座位における遺伝子型データである。この、ハプロタイプとともに浸透率を推定する新しいディプロタイプ形に基づいたアルゴリズムでは、標本空間を定義する試行（または実験）は少し異なる。二つの順番付けのハプロタイプコピーがそれぞれの個体に配られた後、その個体は特定の表現型を非決定論的過程により発現するか発現しない。即ち、ハプロタイプ推定の EM に基づいたアルゴリズムと、ここで発表する新しいアルゴリズムの違いは、新しいアルゴリズムでは表現型の発現という過程が含まれているということである。

1.2.2 新しい標本空間

今、 l 個の連鎖する SNP 座位があるとしよう。すべての可能なハプロタイプの数は $L = 2^l$ である。我々は、無限のハプロタイプコピーの集合を定義する。ここで、ハプロタイプの頻度は $\Theta = (\theta_1, .., \theta_j, .., \theta_L)$ であり、ここで θ_j は j 番目のハプロタイプ頻度である。ただし、$\theta_j \geq 0, \Sigma_{j=1}^{L} \theta_j = 1$ である。N 人の個体のそれぞれに、ハプロタイプコピーの集合よりランダムに引き出して、二つのハプロタイプコピーを順番に与える。$a_1, a_2, .., a_{L^2}$ を可能なディプロタイプ形としよう。i 番目の個体のディプロタイプ形が a_k である確率は $P(d_i = a_k \mid \Theta) = \theta_l \theta_m$ である。ここに d_i は i 番目の個体のディプロタイプ形であり、l, m は a_k を構成するハプロタイプの順番である。これはハプロタイプレベルでのハーディーワインバーグ平衡が仮定されていることを意味している。i 番目の個体は表現型 ψ_+ を d_i の関数で表される確率のもとに発症する。理論的にはすべてのディプロタイプ形に対して浸透率を仮定することが可能である。しかし、すべてのディプロタイプ形に浸透率を対応付けることは現実的ではない。そこで、我々は 2 つの浸透率のみを仮定した。即ち H_{all} をすべてのハプロタイプの集合とし、H_+ を H_{all} の部分集合で、その存在により他と異なった表現型をきたすハプロタイプの集合とする。典型的な例では H_+ はただ一つのハプロタイプを含むが、複数のハプロタイプ

を要素として含むこともできる。もし、H_+ が特定の座位で特定のアレルを含むすべてのハプロタイプの集合と定義すれば、（ハプロタイプではなく）アレルと表現型との関連を検定することと同じになる。

D_+ を H_+ の要素を含むディプロタイプ形の集合としよう。q_+ を i 番目の個体が $d_i \in D_+$ の下で表現型 ψ_+ をきたす確率としよう。そして q_- を i 番目の個体が $d_i \notin D_+$ の条件の下で表現型 ψ_+ を来たす確率とする。

即ち ψ_i を i 番目の個体の表現型とすると、

$$P(\psi_i = \psi_+ \mid d_i \in D_+) = q_+$$

そして

$$P(\psi_i = \psi_+ \mid d_i \notin D_+) = q_-.$$

Θ と q_+, q_- は独立であることに注意。

新しい試行が古い試行と異なるのは、前者では表現型の発生の過程が含められていることである。Θ に加えて q_+ や q_- などのパラメータが確率空間の定義の上で含まれている。

ここで、ψ_i は d_i の条件の下で Θ とは独立であることに注意する。

1.2.3 尤度関数

観察データは個体の遺伝子型、表現型データである。ここで $G_{obs} = (g_1, g_2, .., g_N)$ と $\Psi_{obs} = (w_1, w_2, .., w_N)$ は、それぞれ遺伝子型、表現型のそれぞれの観察データのベクトルとしよう。 ここで g_i と w_i は、それぞれ i 番目の個体の観察される遺伝子型、表現型である。そうすると、尤度関数は次のように成る。

$$L(\Theta, q_+, q_-) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i = a_k \mid \Theta, q_+, q_-) P(\psi_i = w_i \mid d_i = a_k, \Theta, q_+, q_-),$$

ここで A_i は i 番目の個体について g_i に合致する a_k の集合である。

d_i は q_+, q_- と独立であり、ψ_i は d_i の条件下で Θ と独立なので、

$$L(\Theta, q_+, q_-) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i = a_k \mid \Theta) P(\psi_i = w_i \mid d_i = a_k, q_+, q_-), \tag{1}$$

ここで A_i は再び i 番目の個体について g_i に合致する a_k の集合である。

いかなる i と k について、

$$P(\psi_i = w_i \mid d_i = a_k, q_+, q_-) = \begin{cases} q_+ & \text{if } w_i = \psi_+ \text{ and } a_k \in D_+ \\ 1 - q_+ & \text{if } w_i \neq \psi_+ \text{ and } a_k \in D_+ \\ q_- & \text{if } w_i = \psi_+ \text{ and } a_k \notin D_+ \\ 1 - q_- & \text{if } w_i \neq \psi_+ \text{ and } a_k \notin D_+ \end{cases}.$$

表現型が調べている座位に関するディプロタイプ形と独立ならば、尤度関数は

$$L(\Theta, q_0) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(\psi_i = w_i \mid d_i = a_k, q_0) P(d_i = a_k \mid \Theta), \tag{2}$$

ここで q_0 はすべてのディプロタイプ形に対応する浸透率であり、A_i は再び、i 番目の個体について g_i に合致する a_k の集合である。

いかなる i と k に対して、

$$P(\psi_i = w_i \mid d_i = a_k, q_0) = \begin{cases} q_0 & \text{if } w_i = \psi_+ \\ 1 - q_0 & \text{if } w_i \neq \psi_+ \end{cases}.$$

1.2.4 EM アルゴリズム

等式 (1) が Θ, q_+ と q_- の上で最大化され、このようにして得られた最大尤度を L_{max} で表す。そうして、等式 (2) を Θ と q_0 の上で最大化し、そのようにして得られた最大尤度を L_{0max} とする。一般化尤度比

$$L_{0max}/L_{max} \tag{3}$$

をハプロタイプの存在と表現型との関連の検定に用いる。

L_{max} の最大化については、推定すべきパラメータは $\Theta = (\theta_1, \theta_2, ..., \theta_L), q_+$ そして q_- であるが、L_{0max} の最大化については推定すべきパラメータは $\Theta = (\theta_1, \theta_2, ..., \theta_L)$ と q_0 である。後者の最大化において張られる空間は、前者の最大化において張られる空間の部分空間である。帰無仮説の下では $-2\log(L_{0max}/L_{max})$ は自由度 1 の χ^2 分布に従う。

もし $d_1, d_2, ..., d_N$ と $\psi_1, \psi_2, ..., \psi_N$ の完全データが得られるならば、$\theta_1, \theta_2, ..., \theta_L$ と q_+, q_- の最大推定量は $\hat{\theta_j} = n_j/(2N)$、ただし $j = 1, 2, ..., L$、さらに $\hat{q_+} = N_{+\psi+}/N_+$、$\hat{q_-} = N_{-\psi+}/N_-$ のように

簡単に得られる。ただし、n_j は N 人の個体の中の j 番目のハプロタイプのハプロタイプコピーの数である。また、$N_+ = \#\{i; d_i \in D_+\}, N_- = \#\{i; d_i \notin D_+\}, N_{+\psi+} = \#\{i; d_i \in D_+, \psi_i = \psi_+\}$ および $N_{-\psi+} = \#\{i; d_i \notin D_+, \psi_i = \psi_+\}$ である。

ここで $\#\{i;,,\}$ は; の後の条件を満たす個体の数を表す。

しかしながら、完全データは得られず、我々は単に個体の遺伝子型と表現型を観察するのみである。従って、我々は $n_j/(2N)$、$N_{+\psi+}/N_+$ および $N_{-\psi+}/N_-$ の期待値を真の値の変わりに代入する以下の EM アルゴリズムを作成する。

(i) $n = 0$ について初期値 (例えば $\theta_j^{(n)} = 1/L$) を

$\Theta^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)}, .., \theta_L^{(n)})$ に与える、ただし、$\sum_{j=1}^{L} \theta_j^{(n)} = 1$ であり、$\theta_j^{(n)} > 0$ である。

(ii) $n = 0$ について初期値を $q_+^{(n)}, q_-^{(n)}$ に与える。ただし、$0 < q_+^{(n)}, q_-^{(n)} < 1$ である。

(iii) すべての i、そして g_i に合致するすべての a_k について以下を計算する。

$$
\begin{aligned}
& P(d_i = a_k \mid g_i, \psi_i = w_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) \\
&= \quad P(d_i = a_k \mid \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) P(g_i, \psi_i = w_i \mid d_i = a_k, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) \\
&\quad / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) P(g_i, \psi_i = w_i \mid d_i = a_m, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) \qquad (4)
\end{aligned}
$$

ここで、A_i は g_i と合致する a_m の集合である。ここで我々は g_i に合致する a_k のみについて調べることに注意する。さらには d_i は $q_+^{(n)}$、および $q_-^{(n)}$ と独立であり、ψ_i は d_i の条件下で $\Theta^{(n)}$ と独立なので、等式 (4) は以下のようになる。

$$
\begin{aligned}
& P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) = P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) \\
&= \quad P(d_i = a_k \mid \Theta^{(n)}) P(\psi_i = w_i \mid d_i = a_k, q_+^{(n)}, q_-^{(n)}) \\
&\quad / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}) P(\psi_i = w_i \mid d_i = a_m, q_+^{(n)}, q_-^{(n)}) \qquad (5)
\end{aligned}
$$

(iv) N 人の個体に保有されている j 番目のハプロタイプのハプロタイプコピーの数である n_j はランダム変数なので、我々は j 番目のハプロタイプのハプロタイプコピーの数の期待値を定義でき、

$$
\begin{aligned}
& E[n_j | \Psi_{obs}, G_{obs}, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}] \\
&= \quad \sum_{i=1}^{N} \sum_{a_k \in A_i} f_j(a_k) P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)})
\end{aligned}
$$

ここで $f_j(a_k)$ は a_k の中の j 番目のハプロタイプのハプロタイプコピー数であり、A_i は再び、i 番目の個体について g_i に合致する a_k の集合である。ここで $f_j(a_k)$ は 0,1,2 のいずれかである。この期待値をすべての j について計算する。

(v) ここで、$N_{+\psi+}/N_+$ と $N_{-\psi+}/N_-$ はランダム変数であり、従って期待値が定義でき、

$$\begin{aligned} & E[N_{+\psi+}/N_+ \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}] \\ = & \sum_{i=1}^{N} \sum_{a_k \in D_+} y_i P(d_i = a_k \mid \psi_i = w_i, g_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) \\ & / \sum_{i=1}^{N} \sum_{a_k \in D_+} P(d_i = a_k \mid \psi_i = w_i, g_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}), \end{aligned}$$

また

$$\begin{aligned} & E[N_{-\psi+}/N_- \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}] \\ = & \sum_{i=1}^{N} \sum_{a_k \notin D_+} y_i P(d_i = a_k \mid \psi_i = w_i, g_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}) \\ & / \sum_{i=1}^{N} \sum_{a_k \notin D_+} P(d_i = a_k \mid \psi_i = w_i, g_i, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}), \end{aligned}$$

ここで、

$$y_i = \begin{cases} 1 & \text{if } w_i = \psi_+ \\ 0 & \text{if } w_i \neq \psi_+ \end{cases}.$$

(vi) ステップ (iv) の結果から Θ を以下のように、次のステップのために更新する。

$$\theta_j^{(n+1)} = E[n_j | \Psi_{obs}, G_{obs}, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}]/(2N)$$

ステップ (v) の計算結果より、浸透率を次のステップのために以下のように更新する。

$$q_+^{(n+1)} = E[N_{+\psi+}/N_+ \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}]$$

$$q_-^{(n+1)} = E[N_{-\psi+}/N_- \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, q_+^{(n)}, q_-^{(n)}]$$

(vii) (iii) から (vi) までのステップを値が収束するまで繰り返す。

収束した場合の最大推定値を $\hat{\Theta} = (\hat{\theta_1}, \hat{\theta_2}, .., \hat{\theta_L}), \hat{q_+}$ および $\hat{q_-}$ とする。

(viii) 極大を避けるために、さまざまな $\theta_j^{(0)}(j = 1, 2, .., L), q_+^{(0)}, q_-^{(0)}$ の初期値をテストする。

ここで、$P(\Psi_{obs}, G_{obs} \mid \hat{\Theta}, \hat{q_+}, \hat{q_-})$ は、対立仮説の下での最大尤度 L_{max} である。

もし、$q_0 = q_+ = q_-$ の条件を与え、ステップ (iii) から (vii) までを繰り返せば帰無仮説の下での最大尤度 L_{0max} が得られる。

帰無仮説の下では統計量 $-2\log(L_{0max}/L_{max})$ は自由度 1 の χ^2 分布に漸近的に従うと期待される。

欠測データの取り扱い: 観測データの中の欠測データは以下のように取り扱った。即ち、i 番目の個体に遺伝子型データの欠測があった場合は、g_i は g_i と矛盾しない遺伝子型と解釈された。

表現型の欠測データについては、その個体が不明の表現型を来たす確率は 1 と解釈された。

上記のアルゴリズムは、コンピュータソフトウェア PENHAPLO に搭載された [1] [2]。

1.3 ディプロタイプ形と量的表現型の関連の検定とパラメータのアルゴリズム

1.4 標本空間と実験による結果の集合

EM アルゴリズムによる遺伝子型データを用いたハプロタイプと QTL 表現型分布の推定アルゴリズにおいては、標本空間を次のような実験から得られる一つの結果の集合として定義する。連鎖した l 個のＳＮＰ座位に関するハプロタイプがあるとする。可能ハプロタイプの総数は $L = 2^l$ 個である。無数のハプロタイプコピーの集合を定義する。これらのすべてのハプロタイプの頻度は、$\Theta = (\theta_1, ..., \theta_L)$ である。ここで、θ_j は j 番目のハプロタイプの頻度で、$\theta_j \geq 0, \Sigma_{j=1}^{L}\theta_j = 1$ である。次に、N 人の個体にハプロタイプ頻度 Θ に基づき、それぞれ二つずつのハプロタイプを順番に与える。ディプロタイプ形は２つのハプロタイプの順列として定義する。可能なすべての順位付きディプロタイプ形を $a_1, a_2, ..., a_{L^2}$ とする。一つの個体 i が順位付きディプロタイプ形、a_k をもつ確率は $P(d_i = a_k|\Theta) = \theta_l\theta_m$ である。ここで、d_i は個体 i が持つ一つの順位付きディプロタイプ形であり、l,m は a_k を構成するハプロタイプの順位とする。個体 i がＱＴＬ表現型 ψ_i を発現する確率密度関数を考える。ここでは、特定の順位付きディプロタイプ形に対する QTL 表現型は正規分布に従うと仮定する。実験の一つの結果は (Θ, D, Ψ) で表される。$D = (d_1, ..., d_N)$ は個人の順位付きディプロタイプ形のベクトルであり、$\Psi = (\psi_1, ..., \psi_N)$ は個人の表現型のベクトルである。QTL 表現型の分布の平均値 μ は、d_i が他のハプロタイプとは異なる効果をもつ特定のハプロタイプ h_b を含むか含まないかによって決まる。D_+ を特定のハプロタイプ h_b を含むディプロタイプ形の集合とする。ここで、

順位付きディプロタイプ形に対するＱＴＬ表現型の正規分布を２つ定義する。一つは、$N(\mu_1, \sigma^2)$ であり、このとき d_i は特定のハプロタイプを含む。もう一つは、$N(\mu_2, \sigma^2)$ であり、このとき d_i は特定のハプロタイプを含まない。また、ここでの σ^2 は共通の値をとる。$f_{\mu_1}(x)$ は、d_i が特定のハプロタイプを含むとき、個体 i がＱＴＬ表現型 x を発現する確率密度関数である。$f_{\mu_2}(x)$ は、d_i が特定のハプロタイプを含むとき、個体 i がＱＴＬ表現型 x を発現する確率密度関数である。したがって、ψ_i が個体 i のＱＴＬ表現型であるならば、

$$f(\psi_i = x \mid d_i \in D_+) = f_{\mu_1}(x),$$

$$f(\psi_i = x \mid d_i \notin D_+) = f_{\mu_2}(x).$$

ただし、Θ と $f_{\mu_1}(x)$,$f_{\mu_2}(x)$ は独立であり、ψ_i は、d_i の条件下では Θ と独立である。

1.5 尤度関数

観察データは、N 人の個体についての l 個の座位の遺伝子型 g_i とＱＴＬ表現型 w_i についてある。観察データを $G_{obs} = (g_1, g_2, ..., g_N), \Psi_{obs} = (w_1, w_2, ..., w_N)$ とし、それぞれは遺伝子型の観察データと QTL 表現型の観察データを示す。また、$\vec{\mu} = (\mu_1, \mu_2, ..., \mu_{L^2})$ を、可能なすべてのディプロタイプ形のベクトル $A = (a_1, a_2, ..., a_{L^2})$ に対する分布の平均とする。尤度関数は、

$$L(\Theta, \vec{\mu}, \sigma) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i = a_k \mid \Theta, \vec{\mu}, \sigma) f(\psi_i = w_i \mid d_i = a_k, \Theta, \vec{\mu}, \sigma),$$

A_i は g_i に合致する個体 i の順位付きディプロタイプ形 a_k の集合である。

d_i は、$\vec{\mu}, \sigma$ と独立であり、ψ_i は d_i の条件下では Θ とは独立なので、

$$L(\Theta, \vec{\mu}, \sigma) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i = a_k \mid \Theta) f(\psi_i = w_i \mid d_i = a_k, \vec{\mu}, \sigma),$$

A_i は g_i に合致する個体 i のディプロタイプ形の集合である。

個体 i はＱＴＬ表現型 x を次のような確率密度で発現する。

$$f(\psi_i = x | d_i = a_k, \vec{\mu}, \sigma) = \begin{cases} (\frac{1}{\sqrt{2\pi}\sigma}) e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} = f_{\mu_1}(x) & \text{if } a_k \in D_+ \\ (\frac{1}{\sqrt{2\pi}\sigma}) e^{-\frac{(x-\mu_2)^2}{2\sigma^2}} = f_{\mu_2}(x) & \text{if } a_k \notin D_+ \end{cases},$$

ＱＴＬ表現型と順位付きディプロタイプ形は無関係であるという帰無仮説での尤度関数は、

$$L(\Theta, \vec{\mu}, \sigma) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i = a_k \mid \Theta) f(\psi = w_i \mid d_i = a_k, \vec{\mu}, \sigma),$$

ただし、帰無仮説の下での順位付きディプロタイプ形に対するＱＴＬ表現型の分布の平均値は一定であり、これを μ_1、すなわち $\vec{\mu} = (\mu_1, \mu_1, ..., \mu_1)$ である。また、A_i は g_i と合致する個体 i のディプロタイプ形の集合である。

1.6 EM アルゴリズム

式（１）を Θ, $\vec{\mu}$, σ で最大化し、そのとき得られる最大尤度を L_{max} とおく。式（２）を Θ, $\vec{\mu}$, σ で最大化し、そのとき得られる最大尤度を L_{0max} とおく。尤度比 L_{0max}/L_{max} を用いて、ハプロタイプと QTL 表現型の分布に関係があるかどうか検定を行う。

L_{max} において推定すべき変数は、$\Theta = (\theta_1, \theta_2, .., \theta_L)$, $\vec{\mu}$, σ である。また、L_{0max} において推定すべき変数は、$\Theta = (\theta_1, \theta_2, .., \theta_L)$, $\vec{\mu}$, σ である。帰無仮説の下で、$-2log(L_{0max}/L_{max})$ が自由度１の χ^2 分布に従うことを用いて検定を行う。

もし、個体のディプロタイプ形 $d_1, d_2, .., d_N$ と表現型 $\psi_1, \psi_2, .., \psi_N$ についての完全データが利用できるのであれば、$\theta_1, \theta_2, .., \theta_l$, $\vec{\mu}$, σ は $\hat{\theta_j} = n_j/(2N)$ $(j = 1, 2, .., , L)$, $\hat{\mu_1} = \sum_{d_i \in D_+} \psi_i/N_+$, $\hat{\mu_2} = \sum_{d_i \notin D_+} \psi_i/N_-$, $\hat{\sigma} = \sqrt{(\sum_{d_i \in D_+}(\psi_i - \mu_1)^2 + \sum_{d_i \notin D_+}(\psi_i - \mu_2)^2)/n}$ のように簡単に求まる。n_j は N 人の個体における順位付きディプロタイプ形 $d_1, d_2, ..., d_{L^2}$ の中の第 j ハプロタイプの数である。

しかし、完全データを利用できず、今は個体の遺伝子型と QTL 表現型についての観察データしかない。そのため、次のように EM アルゴリズムのなかで真の値、$n_j/(2N)$, $\sum_{d_i \in D_+} \psi_i/N_+$, $\sum_{d_i \notin D_+} \psi_i/N_-$, $\sqrt{(\sum_{d_i \in D_+}(\psi_i - \mu_1)^2 + \sum_{d_i \notin D_+}(\psi_i - \mu_2)^2)/n}$ に対する期待値をおく。

(i)　はじめを $n = 0$ として、初期値、$\Theta^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)}, ..., \theta_L^{(n)})$ を与える。$\sum_{j=1}^{L} \theta_j^{(n)} = 1$、$\theta_j^{(n)} > 0$。

(ii)　はじめを $n = 0$ として、初期値、$\vec{\mu}^{(n)} = (\mu_1^{(n)}, \mu_2^{(n)})$ を与える。

(iii)　はじめを $n = 0$ として、初期値、$\sigma^{(n)}$ を与える。ここでは、σ の値は特定のハプロタイプ h_b

を含むか含まないかにかかわらず同じであると仮定する。

(iv)　すべての個体 i について、それぞれの個体につき g_i と合致するすべての a_k について計算する。

$$
\begin{aligned}
& P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}) \\
= \quad & P(d_i = a_k \mid \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}) f(\psi_i = w_i \mid d_i = a_k, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}) \\
& / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}) f(\psi_i = w_i \mid d_i = a_m, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)})
\end{aligned} \tag{6}
$$

A_i は g_i と合致する a_m の集合である。ここでは、g_i と合致する a_k についてだけ考える。さらに、d_i は $\vec{\mu}^{(n)}, \sigma^{(n)}$ と独立であり、ψ_i は d_i の条件下では $\Theta^{(n)}$ と独立なので、式（3）は、

$$
\begin{aligned}
& P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}) \\
= \quad & P(d_i = a_k \mid \Theta^{(n)}) f(\psi_i = w_i \mid d_i = a_k, \vec{\mu}^{(n)}, \sigma^{(n)}) \\
& / \sum_{a_m \in A_i} P(d_i = a_m \mid \Theta^{(n)}) f(\psi_i = w_i \mid d_i = a_m, \vec{\mu}^{(n)}, \sigma^{(n)})
\end{aligned} \tag{7}
$$

となる。

(v)　N 人の人が持つ j 番ハプロタイプの数 n_j の期待値は確率変数なので、次のように j 番ハプロタイプの期待値を定義する。

$$
\begin{aligned}
& E[n_j \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}] \\
= \quad & \sum_{i=1}^{N} \sum_{a_k \in A_i} g_j(a_k) P(d_i = a_k \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)})
\end{aligned}
$$

$g_j(a_k)$ は a_k の中に含まれている j 番ハプロタイプの数。また、A_i は g_i と合致する個体 i のディプロタイプ形の集合である。ただし、$g_j(a_k)$ は 0,1,2 の値をとり得る。この期待値を全ての j について計算する。

(vi)　$\sum_{d_i \in D_+} \psi_i / N_+$, $\sum_{d_i \notin D_+} \psi_i / N_-$, $\sqrt{(\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \notin D_+} (\psi_i - \mu_2)^2)/n}$ はランダム変数なので期待値が定義でき次のようになる。

$E[\sum_{d_i \in D_+} \psi_i / N_+ \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}] = \frac{\sum_{i=1}^{N} \psi_i (u_b / u_0)}{\sum_{i=1}^{N} (u_b / u_0)}$

$u_b = \sum_{a_k \in D_+} P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, \mu_1^{(n)}, \sigma^{(n)}) f(\psi_i \mid d_i = a_k, \mu_1^{(n)}, \sigma^{(n)})$

$u_0 = \sum_{a_k \in A_i} P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, \mu_1^{(n)}, \sigma^{(n)}) f(\psi_i \mid d_i = a_k, \mu_1^{(n)}, \sigma^{(n)})$

ここで分子と分母は、g_i に合致する個体 i の順位付きデイプロタイプ形の集合のうち表現型に関係したハプロタイプを含む集合の割合、すなわち、u_b/u_0 によってそれぞれ重み付けをしている。

$E[\sum_{d_i \notin D_+} \psi_i/N_+ \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}]$

$= \frac{\sum_{i=1}^{N} \psi_i (v_b/v_0)}{\sum_{i=1}^{N} (v_b/v_0)}$

$v_b = \sum_{a_k \notin D_+} P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, \mu_2^{(n)}, \sigma^{(n)}) f(\psi_i \mid d_i = a_k, \mu_2^{(n)}, \sigma^{(n)})$

$v_0 = \sum_{a_k \in A_i} P(d_i = a_k \mid \psi_i = w_i, \Theta^{(n)}, \mu_2^{(n)}, \sigma^{(n)}) f(\psi_i \mid d_i = a_k, \mu_2^{(n)}, \sigma^{(n)})$

ここで分子と分母は、g_i に合致する個体 i の順位付きデイプロタイプ形の集合のうち表現型に関係したハプロタイプを含まない集合の割合、すわなち v_b/v_0 によってそれぞれ重み付けしている。

$E[\sqrt{(\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \notin D_+} (\psi_i - \mu_2)^2)/N} \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}]$

$= \{(\frac{1}{n} \sum_{i=1}^{N} (\psi_i - \mu_1)^2 \sum_{i=1}^{N} (u_b/u_0) + \frac{1}{n} \sum_{i=1}^{N} (\psi_i - \mu_2)^2 \sum_{i=1}^{N} (v_b/v_0)\}^{1/2}$

ここで σ は、g_i に合致する個体 i の順位付きデイプロタイプ形の集合のうち表現型に関係したハプロタイプを含む集合の割合、すなわち u_b/u_0 と、g_i に合致する個体 i の順位付きデイプロタイプ形の集合のうち表現型に関係したハプロタイプを含まない集合の割合、すなわち v_b/v_0 によってそれぞれ重み付けしている。また、n は $\sum_{i=1}^{N} (u_b/u_0) + \sum_{i=1}^{N} (v_b/v_0)$ とおく。

(vii)　(v) の計算結果より、Θ を次のように更新する。

$$\theta_j^{(n+1)} = E[n_j \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}]/(2N)$$

また、(vi) の計算結果より、$\vec{\mu}$, σ を次のように更新する。

$$\mu_1^{(n+1)} = E[\sum_{d_i \in D_+} \psi_i/N_+ \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}]$$

$$\mu_2^{(n+1)} = E[\sum_{d_i \notin D_+} \psi_i/N_+ \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}]$$

$$\sigma^{(n+1)} = E[\sqrt{(\sum_{d_i \in D_+} (\psi_i - \mu_1)^2 + \sum_{d_i \notin D_+} (\psi_i - \mu_2)^2)/n} \mid \Psi_{obs}, G_{obs}, \Theta^{(n)}, \vec{\mu}^{(n)}, \sigma^{(n)}]$$

(viii)　(iv) から (vii) までをパラメータの値が収束するまで繰り返し実行する。収束したときのパラメータの値が最尤推定値、$\hat{\Theta} = (\hat{\theta_1}, \hat{\theta_2}, .., \hat{\theta_L})$, $\hat{\mu_1}$, $\hat{\mu_2}$, $\hat{\sigma}$ である。

(iX)　Local maximum を避けるため、複数の異なる $\theta_j^{(0)}(j=1,2,..,L^2), \vec{\mu}^{(0)}, \sigma^{(0)}$ について繰り返し計算を行う。ここで、$P(\Psi_{obs}, G_{obs} \mid \hat{\Theta}, \hat{\vec{\mu}}, \hat{\sigma})$ が対立仮説における最大尤度 L_{max} である。また、$\vec{\mu} = (\mu_1, \mu_1, .., \mu_1)$ として、(iv) から (vii) までを繰り返し計算を行うと、帰無仮説におけるを最大尤度 L_{0max} が得られる。帰無仮説の下では、$-2log(L_{0max}(\Theta, \vec{\mu}, \sigma)/L_{max}(\Theta, \vec{\mu}, \sigma))$ が自由度 1 の χ^2 分布に従うことを用いて検定を行う。上記のアルゴリズムは、コンピュータプログラム QTLHAPLO に搭載された [3]。

参考文献

[1] Ito T, Inoue E, Kamatani N. Association test algorithm between a qualitative phenotype and a haplotype or haplotype set using simultaneous estimation of haplotype frequencies, diplotype configurations and diplotype-based penetrances. Genetics. 2004 168:2339-48.

[2] Furihata S, Ito T, Kamatani N. Test of association between haplotypes and phenotypes in case-control studies: examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. Genetics. 2006 174:1505-16.

[3] Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N. Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 2004 168:525-39.

Permutation 法の更なる応用

第 7 章 － 7.6 ハプロタイプ推定のアルゴリズム －

第 8 章 － 8.8 順列並べ換え法の応用 －

	症例	対照	合計
第 1 ハプロタイプ数の期待値	$2n_1p_1$	$2n_2q_1$	$2n_1p_1 + 2n_2q_1$
それ以外のハプロタイプ数の期待値	$2n_1(1-p_1)$	$2n_2(1-q_1)$	$2n_1(1-p_1) + 2n_2(1-q_1)$
合計	$2n_1$	$2n_2$	$2n_1 + 2n_2$

表 1: ハプロタイプ数の期待値の 2×2 表

目 次

1 Permutation 法の更なる応用（第 7 章、7.6)、(第 8 章、8.8)

1.1 ハプロタイプを用いた関連解析に Permutation 法を利用する例

複数の連鎖した SNP 座位の遺伝子型の情報から集団のハプロタイプ頻度を推定する手法については第 7 章、第 5, 6 項で紹介した。このように推定したハプロタイプ頻度を用いて関連解析を行うことができるであろうか。例えば、サイズ n_1 の症例群についてハプロタイプ推定を行い、推定ハプロタイプ頻度 $p_1,\ p_2, \cdots, p_m$ を得た。次に、サイズ n_2 の対照群についてハプロタイプ推定を行い、推定ハプロタイプ頻度 $q_1,\ q_2, \cdots, q_m$ を得た。ここで、m はハプロタイプの総数であり、$p_i,\ q_i$ は i 番目のハプロタイプの頻度を示す。ここで、例えば 1 番目のハプロタイプの頻度に症例と対照で違いがあるかどうかを検定したいとする。症例群における第 1 ハプロタイプ数の期待値は $2n_1p_1$、対照群における期待値は $2n_2q_1$ である。これらの期待値から次の 2×2 表ができる。

表 2 のような偶現表であれば

$$K = \frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{1}$$

	群	
	1	2
属性		
1	a	b
2	c	d

表 2: 2×2 の偶現表

のように帰無仮説の下で χ^2 分布に従うランダム変数 K を計算できる（第 8 章, 第 4 項、式 8.5)。このような計算を観察されたデータを用いて行った場合に得られた K を K_o とする。

しかし、表 1.1 は偶現表ではなく、計算された K が χ^2 分布に従うことは期待できない。ただ、帰無仮説が正しければ（症例と対照で第 1 ハプロタイプの頻度に差がなければ)、このように計算された K の値は小さくなる傾向があることは確かである。帰無仮説から離れるほど K は大きくなる。

ここで、Permutation と呼ばれる操作を行う。まず、症例と対照を併合した $n_1 + n_2$ 人の中からランダムに n_1 人を非復元抽出（一度抽出した人を戻さない）し第一群とする。残りの n_2 人を第二群とする。この二つの群のそれぞれでハプロタイプ頻度推定を行い（第 7 章, 第 6 項)、第 1 ハプロタイプの頻度を、第一、第二群でそれぞれ p_1', q_1' とする。表 1.1 の p_1 を p_1' に、q_1 を q_2' に置き換え、同様に K を式 1 より計算する。このように症例と対照を併合した $n_1 + n_2$ 人の中から n_1 人を抽出する方法は ${}_{n_1+n_2}C_{n_1}$ 通り存在し、それぞれから p_1', q_1' を計算し、K を計算できる。このようにして得られた K を用い、自由度 1 の χ^2 分布を仮定し P 値を計算し、これを P' とする。このように多くの permutation を行うことにより P' の分布のヒストグラムを描く事ができる。これは帰無仮説における P' の分布なので、観察された P_o' がこの分布から著しく大きいほうにずれていれば帰無仮説は否定できる。例えば、P' の帰無仮説の下での分布に照らし合わせ、小さい方の割合 0.05 に入っていれば有意水準 $\alpha = 0.05$ で有意といえる。N_p 個の Permutation 法による P' の値を大きさ順に並び替え小さいほうから i 番目の P' の値を P_i' とし、

$$i/(N_p + 1) \leq \alpha$$

を満足する最大の i を m とし、

$$P'_o \leq P'_m + (P'_{m+1} - P'_m)[\alpha(N_p + 1) - m]$$

なら有意とする。保守的に、$P'_o \leq P'_m$ ならば有意とするという方針でも良いであろう。

このような P' の分布を計算する場合、すべての ${}_{n_1+n_2}C_{n_1}$ 通りについて計算できる場合もある。しかし、計算力の点で困難な場合もあり、その場合は十分の数の Permutation を行い P' を計算した上で帰無仮説の下での分布を計算する。以上は一つのハプロタイプ（ここでは第一ハプロタイプ）のみについてその頻度の違いを検定した。しかし、例えば、第一ハプロタイプと第二ハプロタイプを併合した頻度の違いを検定することも可能である。

しかし、複数のハプロタイプの検定を行うという多重性も考慮に入れる必要がある。その場合は、例えば、頻度 0.05 以上のハプロタイプについてすべて、そのハプロタイプとその他のハプロタイプにわけ、上記の方法で 2×2 の contingency table を用い P' を計算し、最小の P' について Permutation 法で経験分布を調べるという方法もある。重要なことは、それぞれの Permutation について、P' を決める方法が一義的に決まっていることである。特定のハプロタイプをあらかじめ決め、それについてのみ P' を計算することには問題がある。その特定のハプロタイプを決めるという行為が恣意的であるからである。あらかじめ、高い頻度から 2 つ、あるいは 3 つのハプロタイプのみについて P' を計算し、最小のものを採用するという方法もある。

ハプロタイプ頻度だけではなく、ディプロタイプ形の頻度の症例と対照間の違いを検定することも可能である。遺伝統計学入門追加テキスト（Penhaplo と QTLhaplo）で、個人のディプロタイプ形を推定する手法を紹介した。ここで、例えば第一ハプロタイプを保有するディプロタイプ形の集団内の頻度が症例と対照で違うかどうかを検定できる。即ち、遺伝統計学入門追加テキスト（Penhaplo と QTLhaplo）に述べた手法で個人のディプロタイプ形を推定する。個人によってはディプロタイプ形が一つに集中しない場合もあるが、ここでは割り切って、最も頻度の高いディプロタイプ形を選択する。その後、第一ハプロタイプを保有する個体を症例と対照で数え、症例、対照での第一ハプロタイプの保有者の数をそれぞれ a, b、非保有者の数を c, d とする。表 2 のような 2×2 の表を作り、式 1 で P' を計算する。このように観察されたデータから計算された P' を P'_o とする。もちろんこの場合も、特定のハプロタイプではなく、不完全ハプロタイプも含めたすべてのハプロタイプについて

(高頻度のものについてのみでもよい)、保有者、非保有者にわけ P' を計算し、最大の P' を採用しても良い。

P' の帰無仮説の下での分布は上記と類似した Permutation 法を用いる。即ち、表現型をシャッフルし P' の分布を求める。それにより有意水準 α での P' の限界点 P'_m を求め P'_o と比較して検定を行う。

更に、ハプロタイプの保有者と非保有者にわけるのではなく、ハプロタイプの保有数について、0, 1, 2 個の個体を区別し、Cochrane-Armitage 検定を行う方法も考えられる（統計学入門追加テキスト: Cochrane-Armitage)。即ち、Penhaplo ソフトウェアで個人のディプロタイプ形の推定を行った後、あるハプロタイプ（不完全ハプロタイプでもよい）の 0, 1, 2 個の保有者に分ける。個人が一つのディプロタイプ形に定まらない場合は最大確率のディプロタイプ形とする。そのような 2×3 の偶現表を用いてこの場合も、ディプロタイプ形が一義的に決まらない場合も、最大の確率のものを採用すればよい。症例群、対照群についてそのような数を数えた上で 2×3 の偶現表を作成し、その上で Cochrane-Armitage 検定を行う。得られた P 値を保存し、すべてのハプロタイプについて行った最小の P 値を P' とする。この P' の分布から上記と類似の方法により検定を行う。

更に、ディプロタイプ形と質的表現型の関連を検定するアルゴリズム PENHAPLO（遺伝統計学入門追加テキスト: Penhaplo と QTLhaplo)、ディプロタイプ形と量的表現型の関連を検定するアルゴリズム QTLHAPLO（遺伝統計学入門追加テキスト: Penhaplo と QTLhaplo）でも Permutation 法を用い、検定を行うことが可能である。この場合は、上記の P' 統計量ではなく、一般化尤度比を用いると良い。即ち、Permutation により表現型をシャッフルし、それぞれについて一般化尤度比を計算する。そのように帰無仮説の下での分布を求め、観察された一般化尤度比の値と比較する。帰無仮説の下での一般化尤度比の分布より有意に大きいほうに観察された値が偏っていれば有意とする。

1.2 複数のモードで検定を行う場合への FDR の応用

第 8 章, 第 7 項で述べたように、FDR (false discovery rate) による多重比較の補正法は帰無仮説の下での多くの独立した検定の P 値が一様分布をすることを利用した補正法であった。多数の SNP について関連解析を行う場合でも、それぞれの SNP について一つのモード（例えば、allele frequency）

でしか検定を行わない場合は P 値は帰無仮説の下で一様分布に従うことが期待できる。しかし、例えば、多数の SNP について複数のモードで関連解析を行い、それぞれの SNP で最小の P 値を採用する場合がしばしばある（第 8 章, 第 7 項)。そのような場合、P 値が帰無仮説の下で一様分布をすることが期待できない（P 値は最小のものを採用するので、低い方にずれる)。従って、前述の log QQ P-value plot を描いても、$y = x$ の直線には乗らない (図 8-9)。次のように Permutation 法を用いて一様分布に従う P を定義できる。

まず、それぞれの SNP について、P' 値を次のように計算する。

n_1 の症例、n_2 の対照を併合し、$n_1 + n_2$ 人の中からランダムに n_1 人の第一群を抽出する。残りの n_2 人を第二群とし優性モード、劣性モード、アレルモードで関連解析を行い（Pearson's χ^2 法、または Fisher exact 法)、最小の P 値を選択する。このような数多くの Permutation を N_p 回行い、そのような最小の P 値の分布を調べ、N_p 個の中で下から i 番目の最小 P 値を P_i とし、

$$P_i \leq P_o$$

を満足する最大の i を m とし、

$$P' = (m + \frac{P_o - P_m}{P_{m+1} - P_m})/N_p$$

を修正の P 値とする。修正 P 値である P' は帰無仮説の下で一様分布に従うことが期待でき、FDR や log quanile-quantile (QQ) P-value plot (第 8 章, 第 7 項）の作成に用いることができる。

多数の SNP 座位についてこれを行うためには、一回の Permutation についてすべての SNP 座位の最小 P 値を求め、これらを記憶しておき、すべての N_p 回の Permutation が終了した後に各 SNP 座位についての P' を計算するのが良いであろう。

図 8-10 に Permutation 法による P' 値（corrected P value）と、生の P 値（nominal P value）の比較を示す。Permutation 法の方が $y = x$ の直線に近いことを示している。

症例・対照研究の遺伝子型データへの Cochran-Armitage 検定の応用

第 8 章 － 8.5 ３つの遺伝子型を区別した検定法 －

本文第 8 章第 5 節

1　症例・対照研究の遺伝子型データへの Cochran-Armitage 検定の応用（第 8 章，8.5）

1.1　Cochran-Armitage 検定

症例・対照研究から得られた遺伝子型データを用い，2×3 の偶現表を用い，Pearson の χ^2 テスト（自由度 2）を用いて独立性の検定が可能である．しかし，その場合は BB, Bb, bb の遺伝子型の表現型の間の順番を考慮していない．例え，この順番が入れ換わっても検定結果は同じである．しかし，しばしばこれらの遺伝子型の表現型の間には順位がある．質的表現型の場合，ヘテロ接合体の浸透率が二つのホモ接合体の間に来ることが多い．もちろん，完全優性や完全劣性の場合はヘテロ接合体の浸透率はいずれかのホモ接合体と同じになり，超優性の場合はヘテロ接合体の浸透率はホモ接合体の間には来ない．しかし，非常に多くの場合，ヘテロ接合体の浸透率は二つのホモ接合体の浸透率の間に来ると考えられる．これは，遺伝子型の中の特定のアレルの数により表現型，あるいは浸透率が単純増加，あるいは単純減少するという考えである．

このように，遺伝子型の中の特定のアレルの量と表現型の関係が単純増加，あるいは単純減少するという考えの下に，症例・対照研究の遺伝子型データを解析する方法に Cochran-Armitage 検定がある．

もともと，この検定は薬物容量と反応した個体の割合との関連を解析するためにしばしば用いられる．今，r 個の容量，d_1, $d_2, \cdots, d_r$ ただし，$d_1 \leq d_2 \leq d_3 \leq \cdots \leq d_r$ を設定し，それぞれの容量に対応したサンプルサイズ n_j のうち，q_1, $q_2, \cdots, q_r$ の割合が反応すると期待されるとする．ここで，$q_1 \leq q_2 \leq q_2 \leq \cdots \leq q_r$ または，$q_1 \geq q_2 \geq q_2 \geq \cdots \geq q_r$ の単純増加，または単純減少が予想されるとし，それを証明するために実験を行い表 1 の観察データを得た．

表 1 より計算される次の検定統計量（Cochran-Armitage 統計量）

$$\chi_{CA} = \frac{\sum_{j=1}^{r} X_j d_j - \hat{p} \sum_{j=1}^{r} n_j d_j}{\sqrt{\hat{p}(1-\hat{p})[\sum_{j=1}^{n} n_j d_j^2 - \frac{1}{N}(\sum_{j=1}^{r} n_j d_j)^2]}} \tag{1}$$

容量	d_1	d_2	d_3	$\cdots$	d_r
反応ありの人数	X_1	X_2	X_3	$\cdots$	X_r
反応なしの人数	$n_1 - X_1$	$n_2 - X_2$	$n_3 - X_3$	$\cdots$	$n_r - X_r$
各容量のサンプルサイズ	n_1	n_2	n_3	$\cdots$	n_r

表 1:

は，帰無仮説 $q_1 = q_2 = \cdots = q_r$ の下で，標準正規分布に従う．ただし，

$$\begin{aligned} N &= \sum_{j=1}^{r} n_j \\ \hat{p} &= (\sum_{j=1}^{r} X_j)/N. \end{aligned}$$

即ち，χ^2_{CA} は自由度 1 の χ^2 分布に従う．統計量（1）は $d_1 \leq d_2 \leq d_3 \leq \cdots \leq d_r$ であり，$q_1 < q_2 < \cdots < q_r$ の時，分布は正方向に変位し，$q_1 > q_2 > \cdots > q_r$ の時，負方向に変位する．ここで，式（1）は $d'_j = ad_j + b$ のように，すべての d を同じ線形変換しても変化しないことに注意すること．

ここで，3 つの容量の変わりに特定の座位の遺伝子型の中の特定のアレル（a）の数を d_j とし，$AA,\ Aa,\ aa$ の遺伝子型に対し $r = 3,\ d_1 = 0,\ d_2 = 1,\ d_3 = 2$ とする．d_j に対し，反応ありの総人数を $R = \sum_{j=1}^{r} X_j$ とすると，式（1）の二乗は

$$\chi^2_{CA} = \frac{N[N(X_2 + 2X_3) - R(n_2 + 2n_3)]^2}{R(N-R)[N(n_2 + 4n_3) - (n_2 + 2n_3)^2]} \tag{2}$$

ところで，allele frequency モードの関連解析のための Pearson の統計量は

$$\chi^2_{AF} = \frac{2N[N(X_2 + 2X_3) - R(n_2 + 2n_3)]^2}{R(N-R)(n_2 + 2n_3)(2N - n_2 - 2n_3)} \tag{3}$$

従って，

$$\frac{\chi^2_{AF}}{\chi^2_{CA}} = \frac{2[N(n_2 + 4n_3) - (n_2 + 2n_3)^2]}{(n_2 + 2n_3)(2N - n_2 - 2n_3)} \tag{4}$$

遺伝子型 $AA,\ Aa,\ aa$ の個体数 $n_1,\ n_2,\ n_3$ に対し a のアレル頻度を p とすると，HWE のもとではおおよそ，$n_2 = 2p(1-p)N,\ n_3 = p^2N$．これらを式（4）に入れれば，$\chi^2_{AF}/\chi^2_{CA} = 1$．従って，HWE の下では χ^2_{AF} と χ^2_{CA} はほぼ等しい．

さて，純系係数（inbreeding coefficient）は集団から二つのアレルをランダムに選択したとき，それが同祖である確率である．集団のアレル a の頻度を p とし，純系係数を F とすると，遺伝子型頻度はHWEからずれる．一人の持つ二つのアレルが同祖であるとき，それが a である確率は p，また二つのアレルが同祖でないとき，それがどちらも a である確率は p^2 なので，集団における aa の頻度は $Fp+(1-F)p^2$ である．同様に，集団における AA，Aa の確率は $F(1-p)+(1-F)(1-p)^2$，$2(1-F)p(1-p)$ である．

個人における a の数 $N(a)$ の平均は $E[N(a)]=2p$ であるが，分散は

$$\begin{aligned} Var[N(a)] &= [Fp+(1-F)p^2](2-2p)^2+2(1-F)p(1-p)(1-2p)^2+[F(1-p) \\ &+ (1-F)(1-p)^2](0-2p)^2 = 2p(1-p)(1+F) \end{aligned}$$

純系係数を 0 とすると，$Var[N(a)]=2p(1-p)$ であり，近交により分散は $(1+F)$ 倍になる．

G_i，$i=1,\cdots,R$ を i 番目の症例の保有する a アレルの数を示すランダム変数，H_j，$j,\cdots,S$ を対照の保有する a アレルの数を示すランダム変数とし，$T=\sum_i G_i-\sum_j H_j$ とすると，$Var(X\pm Y)=Var(X)+Var(Y)\pm 2Cov(X,Y)$，$Var(X_1+X_2+X_3+\cdots+X_n)=\sum_{i,j}Cov(X_i,X_j)=\sum_{i=1}^n Var(X_i)+\sum_{i\neq j}Cov(X_i,X_j)=\sum_{i,j}Cov(X_i,X_j)=\sum_{i=1}^n Var(X_i)+2\sum_{i<j}Cov(X_i,X_j)$ より

$$\begin{aligned} Var(T) &= Var(\sum_i G_i)+Var(\sum_j H_j)-2Cov(\sum_i G_i,\sum_j H_j) \\ &= \sum_i^R Var(G_i)+2\sum_{i<j}Cov(G_i,G_j)+\sum_{i=1}^S Var(H_i)+2\sum_{i<j}^S Cov(H_i,H_j) \\ &- 2\sum_{i,j}Cov(G_i,G_j) \end{aligned}$$

1.2 症例・対照研究の遺伝子型データへの Cochran-Armitage 検定の応用

症例・対照研究のデータにどのように Cochran-Armitage 検定を適用できるであろうか．

集団の BB，Bb，bb の遺伝子型頻度を p_{BB}，p_{Bb}，p_{bb} とする．それぞれの遺伝子型の浸透率を q_1，q_2，q_3 とする（図 1）．集団において，それぞれの遺伝子型，表現型を持つ個体の頻度（population frequency）は表 2 のようになる．ただし r は罹患率で $r=p_{BB}q_1+p_{Bb}q_2+p_{bb}q_3$ である．

遺伝子型	BB	Bb	bb	合計
症例	$p_{BB}q_1$	$p_{Bb}q_2$	$p_{bb}q_3$	r
対照	$p_{BB}(1-q_1)$	$p_{Bb}(1-q_2)$	$p_{bb}(1-q_3)$	$1-r$
合計	p_{BB}	p_{Bb}	p_{bb}	1

表 2: 集団における各遺伝子型，各表現型の個体の頻度（population frequency）

遺伝子型	BB	Bb	bb	合計
症例内の割合（g_1）	$\frac{p_{BB}q_1}{r}$	$\frac{p_{Bb}q_2}{r}$	$\frac{p_{bb}q_3}{r}$	1
対照内の割合（g_2）	$\frac{p_{BB}(1-q_1)}{1-r}$	$\frac{p_{Bb}(1-q_2)}{1-r}$	$\frac{p_{bb}(1-q_3)}{1-r}$	1

表 3: 症例，対照，それぞれの集団内の各遺伝子型の個体の割合

症例・対照研究の場合は症例群と対照群が別に収集されるので，それぞれの中での相対的割合を考える必要がある．症例群，対照群，それぞれの中での各遺伝子型の相対的割合は，表 3 のようになる．

ここで，問題はそれぞれの遺伝子型について，症例群の中での割合（g_1）と，対照群の中の割合（g_2）の比（g_1/g_2）である．これは各遺伝子型の浸透率が同じであれば $q_1 = q_2 = q_3 = r$ となり 1 となる．

もともと，q_* は浸透率なので，浸透率と遺伝子型との対応を再考する必要がある．

今，遺伝子型 BB, Bb, bb に対応する遺伝子型値を $-a$, d, a とする（「遺伝統計学入門」第 12 章参照）．この座位以外の座位の効果，および環境の効果を加えた表現型値は，それらの遺伝子型値を平均とし，同じ分散 σ^2 を持つ正規分布に従うとする（各座位の遺伝子型値，および表現型値は平均が 0 となるように調製されているとする）（図 2）．このように各座位の遺伝子型値，環境値を加えたものが個人の表現型値となる，というモデルが Fisher による相加的なモデルである．ここで，量的表現型値と浸透率の関係は次のようになる．図 2a で量的表現型が閾値 t 以下であれば発症すると考える．このように閾値以下である確率が浸透率となるので，これは分布を示す量的表現型の確率密度関数を $-\infty$ から t まで積分した値となる（図 2a）．これは標準正規分布の確率密度関数の特定の値（z 変換した値）以下の面積，即ち標準正規分布の累積分布関数に相当する（図 2b）．

正規分布の累積分布関数はロジスティック関数と極めて良く似ている．例えば，標準正規分布の累積分布関数 $y = \Phi(x)$ と $c = 0.61475$ の場合のロジスティック関数

$$y = f_c(x) = \frac{e^{x/c}}{1 + e^{x/c}} = \frac{1}{1 + e^{-x/c}} \tag{5}$$

とはほとんど重なるほど類似している（図 3）．従って，閾値モデルにおける標準正規分布の累積分布関数を，式（5）のロジスティック関数で代用すると，$BB,\ Bb,\ bb$ に対応する浸透率 $q_1,\ q_2,\ q_3$ と，同じ遺伝子型に対応する遺伝子型値 $-a,\ d,\ a$ との関係は，図 4 より

$$\begin{aligned} q_1 &= \Phi(\frac{t+a}{\sigma}) \simeq f_c(\frac{t+a}{\sigma}) \\ q_2 &= \Phi(\frac{t-d}{\sigma}) \simeq f_c(\frac{t-d}{\sigma}) \\ q_3 &= \Phi(\frac{t-a}{\sigma}) \simeq f_c(\frac{t-a}{\sigma}) \end{aligned} \tag{6}$$

となる．

式（5）の逆関数は

$$y = c \log \frac{x}{1-x} \tag{7}$$

であり，$y = \Phi^{-1}(x)$ とほぼ一致するので，式（6）より

$$\begin{aligned} \frac{t+a}{\sigma} &= \Phi^{-1}(q_1) \simeq c \log \frac{q_1}{1-q_1} \\ \frac{t-d}{\sigma} &= \Phi^{-1}(q_2) \simeq c \log \frac{q_2}{1-q_2} \\ \frac{t-a}{\sigma} &= \Phi^{-1}(q_3) \simeq c \log \frac{q_3}{1-q_3} \end{aligned} \tag{8}$$

そして，量的形質がアレルに関して相加的という意味は，

$d = (-a + a)/2 = 0$ という事であり，

$$\frac{t-d}{\sigma} = (\frac{t+a}{\sigma} + \frac{t-a}{\sigma})/2$$

ということであるが，これは式（8）より，

$$\log \frac{q_2}{1-q_2} = (\log \frac{q_1}{1-q_1} + \log \frac{q_3}{1-q_3})/2$$

遺伝子型	BB	Bb	bb	合計
症例 (g_1)	$p_{BB}q_1/r$	$p_{Bb}q_2/r$	$p_{bb}q_3/r$	1
対照 (g_2)	$p_{BB}(1-q_1)/(1-r)$	$p_{Bb}(1-q_2)/(1-r)$	$p_{bb}(1-q_3)/(1-r)$	1
相対的割合の比 (g_1/g_2)	$s_1 = \frac{q_1}{1-q_1}/\frac{r}{1-r}$	$s_2 = \frac{q_2}{1-q_2}/\frac{r}{1-r}$	$s_3 = \frac{q_3}{1-q_3}/\frac{r}{1-r}$	

表 4: 症例・対照のそれぞれの群における各遺伝子型の相対的割合とその比

ということである．即ち，量的形質に関してアレルの効果の相加性を仮定するということは，質的形質については，それぞれの遺伝子型での疾患のあるなしのオッズ（即ち $q*/(1-q*)$）の対数がアレルの数に関して線形である事を意味している．

またこれは，表 4 において，症例・対照研究の 2 つのホモ接合体の相対的割合の比（s_1, s_3）の対数の平均が，ヘテロ接合体の相対的割合の比（s_2）の対数に等しくなる事を示している（$\log s_2 = (\log s_1 + \log s_3)/2$）．

症例・対照研究において実際に得られる標本は，症例・対照の標本サイズが m_1, m_2 の時，表 5 のようなものである．

ここで，相対的割合の比（g_1/g_2）の推定量は，表 5 の通りである．即ち，これらの推定量を用いてアレルの効果の相加性を検討することが出来る．

即ち，次の統計量

$$D = \frac{m_2}{m_1}[\frac{X_2}{Y_2} - \frac{1}{2}(\frac{X_1}{Y_1} + \frac{X_3}{Y_3})]$$

が 0 に近いほど，アレルの効果が相加的である事を意味し，$q_1 > q_3$ の場合，(+) に偏れば Bb の表現型が相加的な仮定よりも BB に近く，(-) に偏れば bb に近いことをしめす．$q_1 < q_3$ の場合は逆の関係である．

もし，表現型と遺伝子型の間に全く関連が無ければ，各遺伝子型について，相対的割合の比 g_1/g_2 は同じになるはずである．即ち，表 4, 表 5 より，ランダム変数

$$\frac{X_1}{Y_1}, \frac{X_2}{Y_2}, \frac{X_3}{Y_3}$$

は同じパラメータ（m_1/m_2）の推定量となる．

遺伝子型	BB	Bb	bb	標本サイズ
症例	X_1	X_2	X_3	m_1
対照	Y_1	Y_2	Y_3	m_2
合計	n_1	n_2	n_3	N
相対的割合の比の推定量	$\frac{m_2X_1}{m_1Y_1}$	$\frac{m_2X_2}{m_1Y_2}$	$\frac{m_2X_3}{m_1Y_3}$	

表 5:

ここで，例えばアレル B が浸透率を高める作用があるとすると，$q_1 \geq q_2 \geq q_3$ となり，相対的割合の比は $s_1 \geq s_2 \geq s_3$ となる．

アレル B が浸透率を低める作用があるとすると，逆に $q_1 \leq q_2 \leq q_3$ となり，$s_1 \leq s_2 \leq s_3$ となるであろう．

以上の考察により，表 5 のような症例・対照研究から得られる遺伝子型データについて，次のような統計量を計算する．

$$\chi'_{CA} = \frac{X_3 - X_1 - \hat{p}(n_3 - n_1)}{\sqrt{\hat{p}(1-\hat{p})[n_1 + n_3 - (n_1 - n_3)^2/(m_1 + m_2)]}} \tag{9}$$

ただし，$\hat{p} = m_1/(m_1 + m_2)$. ここで式（1）において，$r = 3,\, d_1 = -1,\, d_2 = 0,\, d_3 = 1$ としたことに注意．d_j については，b アレルの量である $d_1 = 0,\, d_2 = 1,\, d_3 = 2$ でもいいが，線形変換を行っても χ_{CA} 統計量は変化しない．

χ'^2_{CA} は帰無仮説の下（この座位と表現型に関連が無い）では自由度 1 の χ^2 分布に従うので，$\chi'^2_{CA} \geq \chi^2_{1-\alpha}$ であれば有意とする，ただし，α は有意水準で，χ^2_β は自由度 1 の χ^2 分布の β における累積分布関数の値である．もちろん片側検定としたければ，χ_{CA} が標準正規分布に従うことを利用して，増加，減少の傾向に応じ，それぞれ $\chi_{CA} \geq z_{1-\alpha}$ または，$\chi_{CA} \leq z_\alpha$ の基準を使えばよい．ただし z_β は標準正規分布の累積分布関数が β となる点の値である．

1.3　オプション: 症例・対照研究における，アレルの効果の相加性の検定

図 4，および式（6）より，それぞれの遺伝子型の個人が症例となる確率は浸透率 $q_1,\ q_2,\ q_3$ であり，$a,\ d,\ t, \sigma$ を用いて標準正規分布の累積分布関数（Φ）またはロジスティック関数で表される．ここで，式（6）は $t' = t/\sigma,\ a' = a/\sigma,\ d'\ d/\sigma$ と置いて，

$$
\begin{aligned}
q_1 &= \Phi(t' + a') \simeq f_c(t' + a') = 1/(1 + e^{-(t'/c + a'/c)}) \\
q_2 &= \Phi(t' - d') \simeq f_c(t' - d') = 1/(1 + e^{-(t'/c - d'/c)}) \\
q_3 &= \Phi(t' - a') \simeq f_c(t' - a') = 1/(1 + e^{-(t'/c - a'/c)})
\end{aligned}
\tag{10}
$$

と，$t',\ a',\ d'$ の 3 つのパラメータで表される．

表 5 のようなサンプルが得られたとき，対数尤度関数は（ここで、データが集団からのランダムなサンプルではなく、疾患あり（確率 r）、疾患なし（確率 $1-r$）でまず選択されていることに注意が必要である。従って、症例一人の遺伝子型の確率は、症例の中の割合になるが、r は罹患率であり、

$$
r = p_{BB}q_1 + p_{Bb}q_2 + p_{bb}q_3
$$

であり、HWE を仮定すれば

$$
r = p^2 q_1 + 2p(1-p)q_2 + (1-p)q_3 \tag{11}
$$

であり、対数尤度は C を定数とし、

$$
\begin{aligned}
l(t', a', d', p) &= X_1(\log p^2 q_1 - \log r) + Y_1[\log p^2(1-q_1) - \log r] + X_2[\log 2p(1-p)q_2 - \log r] \\
&+ Y_2[\log 2p(1-p)(1-q_2) - \log(1-r)] + X_3[\log(1-p)^2 q_3 - \log(1-r)] \\
&+ Y_3[\log(1-p)^2(1-q_3) - \log(1-r)] + C
\end{aligned}
\tag{12}
$$

これに，式（11）より r を代入すると、式（12）は p, q_1, q_2, q_3 の式となるが、q_1, q_2, q_3 に、式（10）の f_c，または Φ の式を代入すれば，l は $t',\ a',\ d',\ p$ の関数となる．これを最大化する $\hat{t}',\ \hat{a}',\ \hat{d}',\ \hat{p}$ を求め，この時の最大対数尤度を l_{max} とする．また，$d' = 0$ と置いて，式（12）を $t',\ a'$ で最大化し，この時の最大対数尤度を l_{0max} とする．

次の統計量

$$-\frac{1}{2}(l_{0max} - l_{max})$$

は帰無仮説の下で自由度 1（パラメータの数の差）の χ^2 分布に従うと予想される．これを用いて，質的表現型のアレルの効果の相加性の検定を，症例・対照研究のサンプルを用いて行うことが可能である．ただし、パラメータの数が多いので、この手法でパラメータの推定や $d = 0$ の検定が実際に出来るかわからない。p は症例と対照の、アレル B の割合と r から計算し、与えるという方法もあるであろう。

ここで，式（10）の f_c を用いた場合の対数尤度関数（12）の式は，ロジスティック回帰の対数尤度関数と等しい．

即ち，次のロジスティック関数

$$y = \frac{1}{1 + e^{-(\alpha_1 x_1 + \alpha_2 x_2 + \beta)}} \tag{13}$$

あるいは、ロジット関数

$$\log(\frac{y}{1-y}) = \alpha_1 x_1 + \alpha_2 x_2 + \beta$$

を考え、BB の場合は $(x_1 = -1,\ x_2 = 0)$，Bb の場合は $(x_1 = 0,\ x_2 = 1)$，bb の場合は $(x_1 = 1,\ x_2 = 0)$ として、ロジスティック回帰で $\alpha = 0$ の検定を行えばよい．ただし，$\alpha_1 = -a/(c\sigma),\ \alpha_2 = d/(c\sigma),\ \beta = t/(c\sigma)$ である．

1.4 Cochran-Armitage 検定の検出力

唯一の疾患関連座位に二つのアレル $B,\ b$ が存在し，B の集団内頻度を p とする．

遺伝子型 $BB,\ Bb,\ bb$ の遺伝子型値を $-a,\ d,\ a$ とし，非遺伝的分散を σ^2 とする．それぞれの遺伝子型の表現型値の分布は，平均 $-a,\ d,\ a$，分散 σ^2 の正規分布に従うとする．表現型値が特定の値（閾値）以上の場合に発症するとすると，それぞれの遺伝子型に対応する浸透率は表 6 のような標準正規分布の累積分布関数により表される（図 1, 2 を参照）．

表 6 から，罹患率

$$r_+ = p^2\Phi(\frac{t+a}{\sigma}) + 2p(1-p)\Phi(\frac{t-d}{\sigma}) + (1-p)^2\Phi(\frac{t-a}{\sigma}) \tag{14}$$

遺伝子型	BB	Bb	bb
集団内遺伝子型頻度	p^2	$2p(1-p)$	$(1-p)^2$
浸透率	$\Phi(\frac{t+a}{\sigma})$	$\Phi(\frac{t-d}{\sigma})$	$\Phi(\frac{t-a}{\sigma})$
患者（集団内頻度）	$p^2\Phi(\frac{t+a}{\sigma})$	$2p(1-p)\Phi(\frac{t-d}{\sigma})$	$(1-p)^2\Phi(\frac{t-a}{\sigma})$
非患者（集団内頻度）	$p^2[1-\Phi(\frac{t+a}{\sigma})]$	$2p(1-p)[1-\Phi(\frac{t-d}{\sigma})]$	$(1-p)^2[1-\Phi(\frac{t-a}{\sigma})]$

表 6: 患者，非患者中の各遺伝子型の集団内頻度（ただし Φ は標準正規分布の累積分布関数）

非罹患率

$$r_- = p^2[1-\Phi(\frac{t+a}{\sigma})] + 2p(1-p)[1-\Phi(\frac{t-d}{\sigma})] + (1-p)^2[1-\Phi(\frac{t-a}{\sigma})] \tag{15}$$

が計算できる．

ここで，症例の総数（m_1）と対照の総数（m_2）が与えられれば，患者群に対し，次の頻度パラメータを用い，

$$(p^2\Phi(\frac{t+a}{\sigma})/r_+,\ 2p(1-p)\Phi(\frac{t-d}{\sigma})/r_+,\ (1-p)^2\Phi(\frac{t-a}{\sigma})/r_+)$$

m_1 より三項分布を用いて，症例の中の各遺伝子型の個体の数（$X_1,\ X_2,\ X_3$）を得ることが出来る．

対照群に対し，次の頻度パラメータを用い

$$(p^2[1-\Phi(\frac{t+a}{\sigma})]/r_-,\ 2p(1-p)[1-\Phi(\frac{t-d}{\sigma})]/r_-,\ (1-p)^2[1-\Phi(\frac{t-a}{\sigma})]/r_-)$$

m_2 より三項分布を用いて対照群の中の各遺伝子型の個体の数（$Y_1,\ Y_2,\ Y_3$）の標本を得ることが可能である．

即ち，表 6 のような偶現表が出来る．

これらのデータを用い，アレル頻度を用いた関連解析，Cochran-Armitage 検定を用いた関連解析は容易に出来る．従って，標本の生成，Cochran-Armitage 検定を用いた関連解析を繰り返し有意となる頻度を計算することにより検出力を推定できる．

遺伝子型	BB	Bb	bb	合計
		相乗的効果を仮定しない場合		
症例内の割合	$\frac{p^2q}{r}$	$\frac{2p(1-p)qs}{r}$	$\frac{(1-p)^2qt}{r}$	1
対照内の割合	$\frac{p^2(1-q)}{1-r}$	$\frac{2p(1-p)(1-qs)}{1-r}$	$\frac{(1-p)^2(1-qt)}{1-r}$	1
		相乗的効果の場合		
症例内の割合	$\frac{p^2q}{r'}$	$\frac{2p(1-p)qs}{r'}$	$\frac{(1-p)^2qs^2}{r'}$	1
対照内の割合	$\frac{p^2(1-q)}{1-r'}$	$\frac{2p(1-p)(1-qs)}{1-r'}$	$\frac{(1-p)^2(1-qs^2)}{1-r'}$	1

表 7: 浸透率に対するアレル数の相乗性を仮定した場合の症例，対照，それぞれの集団内の各遺伝子型の個体の割合

1.5 浸透率に対するアレルの数の相乗性を仮定した場合

アレル B の頻度を p、遺伝子型 BB の個体の浸透率を q、遺伝子型 $Bb,\ bb$ の個体の浸透率をそれぞれ $qs,\ qst$ とする。即ち、Bb と BB の浸透率の比は s、bb と Bb の浸透率の比は t である。ここで、$s = t$ の時、浸透率が相乗的なモデルとなる。

HWE の仮定の下、浸透率 r は

$$r = p^2q + 2p(1-p)qs + (1-p)^2qst$$

この r を用いて、症例・対照研究において、症例内、対照内の各遺伝子型の個体の割合の期待値は表 7 のようになる。ただし、r' は相乗的効果を仮定した場合の r であり、以下の式で表される。

$$r' = p^2q + 2p(1-p)qs + (1-p)^2qs^2$$

表 5 のような症例対照研究のデータが得られたとすると、その対数尤度は相乗的性を仮定しない

場合

$$\begin{aligned} l(p,q,s,t) &= X_1(\log p^2q - \log r) + Y_1[\log p^2(1-q) - \log r] + X_2[\log 2p(1-p)qs - \log r] \\ &+ Y_2[\log 2p(1-p)(1-qs) - \log(1-r)] + X_3[\log(1-p)^2qst - \log(1-r)] \\ &+ Y_3[\log(1-p)^2(1-qst) - \log(1-r)] + C \end{aligned} \tag{16}$$

ただし C は定数である。

相乗性を仮定する場合

$$\begin{aligned} l_0(p,q,s) &= X_1(\log p^2q - \log r') + Y_1[\log p^2(1-q) - \log r'] + X_2[\log 2p(1-p)qs - \log r'] \\ &+ Y_2[\log 2p(1-p)(1-qs) - \log(1-r')] + X_3[\log(1-p)^2qs^2 - \log(1-r')] \\ &+ Y_3[\log(1-p)^2(1-qs^2) - \log(1-r')] + C \end{aligned} \tag{17}$$

ただし C は定数である。

(i) 表 5 の観察データの下に、式 (16) を p,q,s,t の上に最大化し、最大尤度を $\hat{l}$ とする（この時、$0 < p < 1,\ q > 0,\ s > 1,\ t > 1,\ r = p^2q + 2p(1-p)qs + (1-p)^2qst < 1$ の制限を用いる必要がある)。これが相乗性を仮定しない場合の（4 つのパラメータの下での）最大尤度である。

次に、同じ表 5 の観察データの下に、式（17）を p,q,s の上に最大化し、最大尤度を $\hat{l_0}$ とする（この時、$0 < p < 1,\ q > 0,\ s > 1,\ r = p^2q + 2p(1-p)qs + (1-p)^2qs^2 < 1$ の制限を用いる必要がある)。

$-2(\hat{l} - \hat{l_0})$ は帰無仮説の下で（実際には相乗的効果が完全に成り立つ場合)、自由度 1 の χ^2 分布に従う。

(ii) ただし、最尤推定における q の推定はかなり悪いようである。ここで、q を別のデータから定数として与える方法がある。

q を定数として与えれば、式 (16) は p,s,t の 3 変数、式 (17) は p,s の 2 変数の式となる。その場合は、$\hat{l}$ は 3 変数の上での最大尤度、$\hat{l_0}$ は 2 変数の上での最大尤度となり、$-2(\hat{l} - \hat{l_0})$ は帰無仮説の下で（実際には相乗的効果が完全に成り立つ場合)、自由度 1 の χ^2 分布に従う。

図1

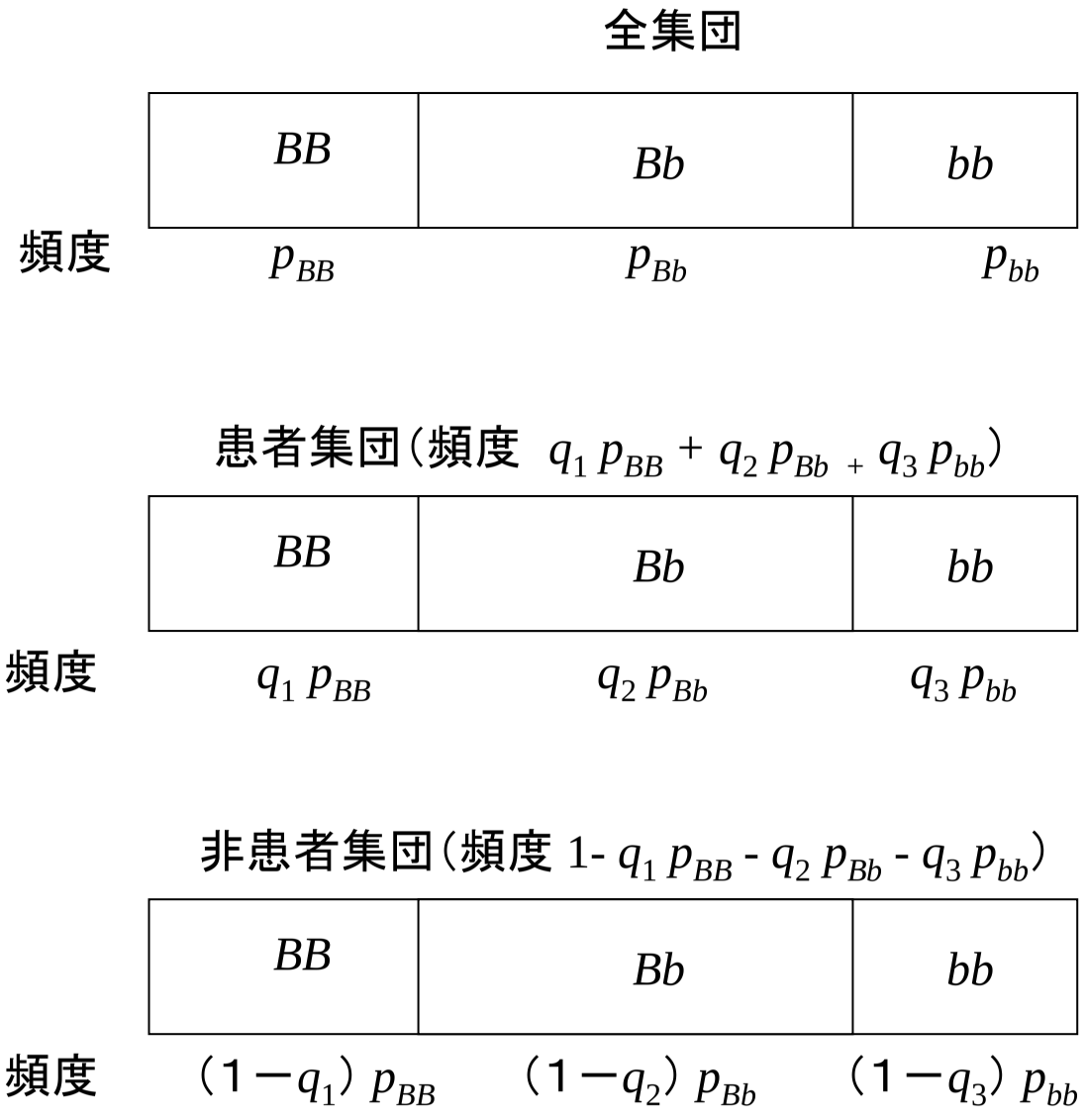
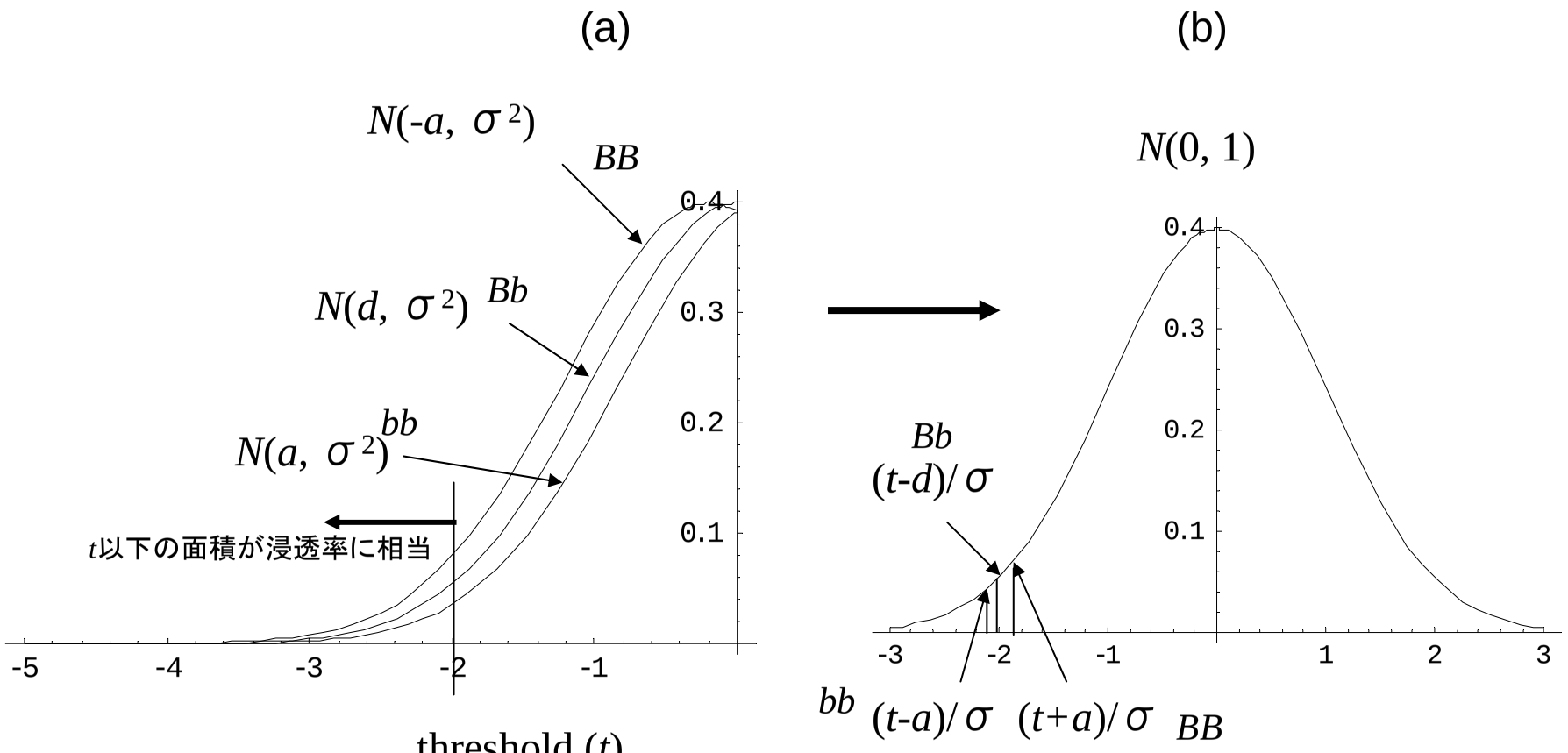

全集団
BB
Bb
bb
頻度
p_{BB}
p_{Bb}
p_{bb}
患者集団（頻度 $q_1 p_{BB} + q_2 p_{Bb} + q_3 p_{bb}$）
BB
Bb
bb
頻度
$q_1 p_{BB}$
$q_2 p_{Bb}$
$q_3 p_{bb}$
非患者集団（頻度 $1- q_1 p_{BB} - q_2 p_{Bb} - q_3 p_{bb}$）
BB
Bb
bb
頻度
$(1-q_1) p_{BB}$
$(1-q_2) p_{Bb}$
$(1-q_3) p_{bb}$

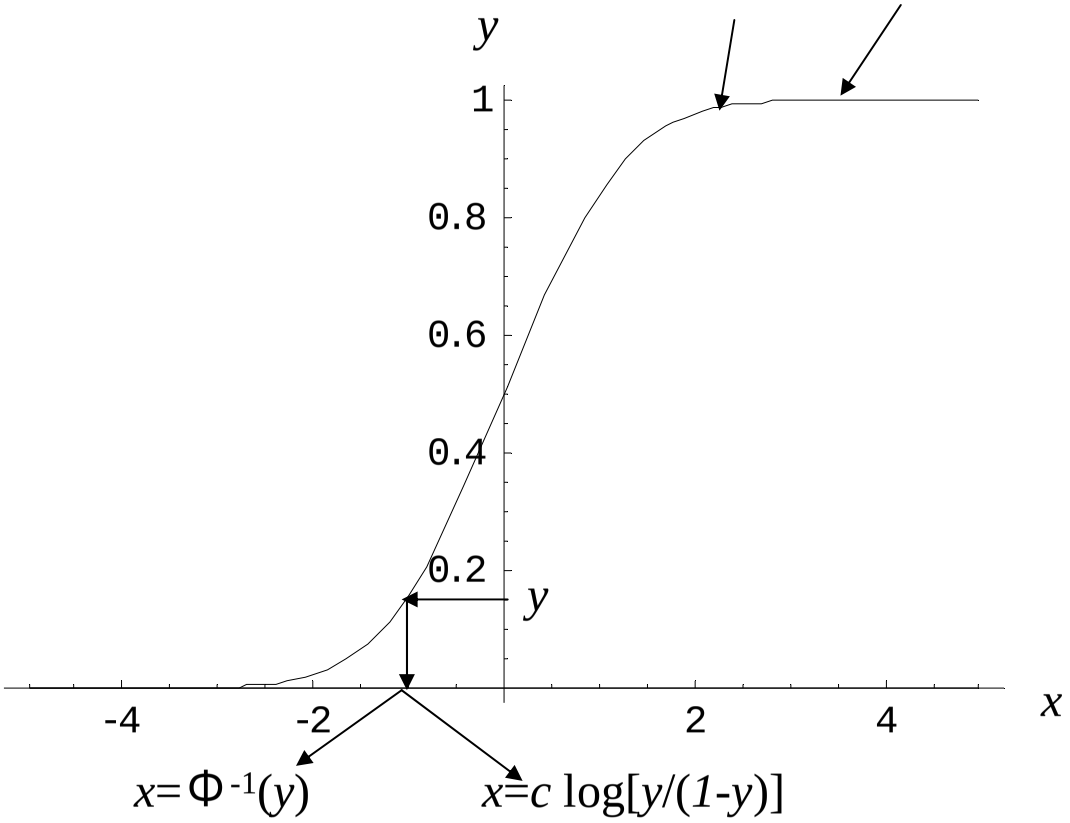
図 2　量的表現型値の分布と浸透率の関係



標準正規分布の確率密度関数における特定の値（z変換した値）以下の面積、即ち、その特定の値における累積分布関数に相当

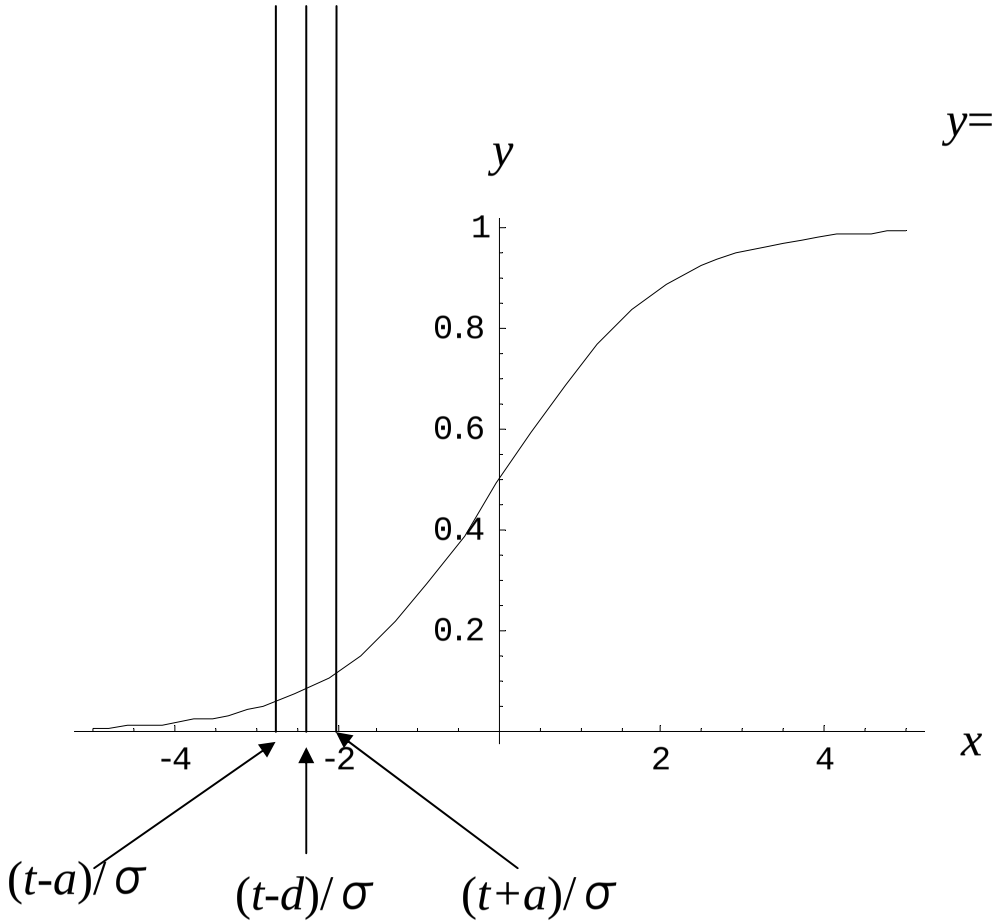
図3

$N(0, 1)$の累積分布関数 $y=\Phi(x)$,　　$y=1/(1+e^{-x/c})$;　$c=0.61475$



ロジスティック関数（c=0.61475の時）と平均0、標準偏差1の正規分布の
累積分布関数は極めて良く似ている（重なっているので違いがわからない）

図4



$y=1/(1+e^{-x/a})$

逆関数

$x=a\ \log[y/(1-y)]$

アレルが相加的に作用する場合、ロジスティック関数を利用し、浸透率（y）を使ってオッズ $y/(1-y)$ を計算すると、*Bb*に対応するオッズの対数は*BB, bb*の平均になる

関連解析の有意性の閾値設定のためのベイズ的方法

第 8 章 － 8.7 網羅的 SNP 関連解析における多重検定の問題 －

真実	検定の結果		
	有意でないと出る	有意と出る	計
関連なし	$1-\alpha$	α	1
関連あり	β	1-β	1

表 1: 関連なし、関連ありのそれぞれの場合の検定結果の確率（条件付確率）

1 関連解析の有意性の閾値設定のためのベイズ的方法（第８章、8.7）

多数の全く情報の無い SNP を用いた網羅的関連解析で見つかった疾患関連 SNP と、もともと疾患と関連がありそうな遺伝子（候補遺伝子）の SNP が同じ P 値を示した場合、どちらが本当の可能性が高いであろうか。やはり、候補遺伝子の SNP の方が本当の可能性が高いであろう。なぜなら、もともと関連がありそうな遺伝子の SNP は統計的検定前から関連がある可能性が高いからである。この要素を取り入れるためには事前確率（prior または prior probability）の概念を取り入れればよい。

ある SNP が、ある形質に本当に関連している確率を π とする（事前確率）。関連していない確率は $1-\pi$ なので、関連しているか、関連していないかを示すオッズ（関連している確率/関連していない確率）は $\pi/(1-\pi)$ である。

ここで、タイプ I の過誤率を α、タイプ II の過誤率が β（検出力は $1-\beta$）の検定を行ったとする。つまり、本当には関連していなくても α の確率で関連ありと判定され、本当には関連していても β の確率で関連なしと判定される（表 1）。つまり、これは、「本当に関連がある」、「本当は関連が無い」という**条件**のもとに、その SNP が検定で「関連あり」、「関連なし」と判定される確率、即ち、「条件付確率」を示している。

ある第一の出来事（ここでは本当に関連がある）の事前確率（π）と、その出来事の条件で別の第二の出来事（ここでは検定結果が有意）が起きる確率（$1-\beta$）を乗じたものが、第一と、第二の出来事がどちらも起きる確率、$(1-\beta)\pi$ である。

同様に、事前の出来事として、「真に関連あり」「真に関連なし」、条件付の出来事として「検定が有意と出る」「有意でないと出る」という設定で、どちらも起きる確率を計算すると、表 2 のようになる。

真実	検定の結果		
	有意でないと出る	有意と出る	計
関連なし	$(1-\alpha)(1-\pi)$	$\alpha(1-\pi)$	1-π
関連あり	$\beta\pi$	$(1-\beta)\pi$	π
計	$(1-\alpha)(1-\pi)+\beta\pi$	$\alpha(1-\pi)+(1-\beta)\pi$	

表 2: 事前確率を考慮した検定結果の確率

ここで、「真に関係ある」:「真に関係ない」のオッズを考える。このオッズが高いほど真に関係がある可能性が高くなる。このオッズは主観的にこの SNP が真に関係しているかどうかを判断する基準であり、検定を行う前と行った後で変化しても良い。このあたりがベイズ推定の正当性に関する議論があるところである。もともと、関係あるか無いかは確率 1 か 0 であり、それ以外はありえないので、このようなオッズで真実の確率を検討することは無意味とする考えもありうる。

しかし、ここでこの SNP が「真に関係ある」:「真に関係ない」のオッズは、検定前には $\pi/(1-\pi)$ であった。しかし、検定結果が「有意である」、と出た場合、このオッズは $(1-\beta)\pi/[\alpha(1-\pi)] = \pi/(1-\pi)\times(1-\beta)/\alpha$ となった（表 2)。つまり、検定で有意と出たことによりオッズは $(1-\beta)/\alpha$ 倍になったのである。

つまり、

真の関連を示す事後のオッズ = 真の関連を示す事前のオッズ × 検出力 / 有意水準

ということになる。

ここで重要なことは、検定による真実性の程度は、その検定の有意水準だけではなく、検出力が関係しているということである。

ベイズの定理を用いて、表 2 より、検定で有意と出たという条件の下で、真実は関連の無い確率（FPRP; false-positive report probability）（事後確率）を計算すると [**?**]

$$R = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi} \tag{1}$$

となる。

ただし、α は検定で勝手に設定するので問題ないが、β と π の計算は容易ではない。

$1-\beta$、即ち検出力の計算には、これから見つけようとする SNP の効果の大きさ（effect size）を仮定する必要がある。即ち、症例・対照研究などでは症例と対照群の間のアレル頻度や遺伝子型頻度に関するオッズ比などである（ここで言う、効果サイズのオッズ比は、関連の真実性を示すオッズとは全く異なるので注意)。また、π を代入するためにはかなりの主観が入らざるを得ないであろうが、例えば、今調べようとしている 50 万 SNP の中の 10 個位は真に関連しているであろう、というような議論から $\pi = 2 \times 10^{-5}$ などと推定する。これに対し、有意水準 $\alpha = 5 \times 10^{-7}$ で、検出力 $1-\beta = 0.5$ の検定により「有意」という結果が出たとすると、その SNP の真実性を示すオッズは $(1-\beta)/\alpha = 10^6$ 倍になったのである。即ち、検定が有意と出ることにより真実性を示すオッズは $2 \times 10^{-5} \times 10^6 = 20$ となった。即ち、有意と出た SNP が真に関連する可能性は関連しない可能性の 20 倍である。FPRP は (1/(20+1)=0.048 である。

参考文献

[1] Ito T, Inoue E, Kamatani N. Association test algorithm between a qualitative phenotype and a haplotype or haplotype set using simultaneous estimation of haplotype frequencies, diplotype configurations and diplotype-based penetrances. Genetics. 2004 168:2339-48.

[2] Furihata S, Ito T, Kamatani N. Test of association between haplotypes and phenotypes in case-control studies: examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. Genetics. 2006 174:1505-16.

[3] Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N. Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 2004 168:525-39.

RAT: rapid association test

第 8 章 － 8.8 順列並べ換え法の応用 －

1 RAT: rapid association test（第8章、8.8）

1.1 標本空間

一つの実験:

ξ 人の患者、$n-\xi$ 人のコントロールについて関連解析を行う。すべての個人について m 座位の遺伝子型が観察されているとする。ここで、$1,2,\cdots,\xi$ 番目の患者、$\xi+1,\xi+2,\cdots,n$ 番目のコントロールについて permutation を行う。新たに $1,2,\cdots,\xi$ 番目の患者、$\xi+1,\xi+2,\cdots,n$ 番目のコントロールの場所を設け、n 人の個人からランダムに順番に選択する。これにより、新たなケース・コントロール集団ができる。このような permutation を一つの実験とする。

標本空間:

以上の一つの実験による結果を ω とする。すべての ω の集合を標本空間 Ω とする。ω の数は $n!$ 個ある（ω を permutation outcome という事にする）。選択がランダムであれば、ω はすべて等確率、$1/n!$ である（確率測度）。

出来事:

異なった ω であっても、ケース・コントロール集団における個人の順番は異なっても組合せは全く同じものがある。ケース・コントロールで組合せが同じ ω の数は、それぞれの中での順番の入れ換え方の数の積であり $\xi!(n-\xi)!$ 個ある。そのような、それぞれの群で個人の組合せが同じ ω を集めた Ω の部分集合を λ（これを permutation event という事にする）とすると Ω は

$$n!/(\xi!(n-\xi)!) = {}_nC_\xi$$

個の等確率の出来事に分割（partition）される。一つの λ の確率は $1/{}_nC_\xi$ である。

1.2 Permutation event

$\mathbf{d}_i$ を ξ 個の要素 1、$n-\xi$ 個の要素 0 を持つ n 次元のベクトルとすると、一つの出来事 λ は一つの $\mathbf{d}_i$ と対応する。これは n 人の個人の ξ 人に疾患のラベルを貼り、$n-\xi$ 人に非疾患のラベルを貼る事であり、これは ξ 人の疾患群と $n-\xi$ 人の非疾患群を集めることと同じことが理解できるであろう。

Status	genotype 1	genotype 2	total
disease	$T_{0,0}$	$T_{1,0}$	ξ
control	$T_{0,1}$	$T_{1,1}$	$n-\xi$
total	n_1	n_2	

表 1: 偶現表 T

従って、$\mathbf{d}_i$ を一つの permutation event と考えることができ、その確率は

$$Pr(\mathbf{d}_i) = \frac{1}{{}_nC_\xi}$$

一つのマーカー座位 j について、一つの ω が決まると一つの偶現表ができるが、同じ λ の要素である ω によって出来る偶現表は皆同じである。しかし、j についての一つの偶現表は多数の λ に（従って、多数の $\mathbf{d}_i$ に）対応し、偶現表 T（表 1）に対応する λ の数（従って $\mathbf{d}_i$ の数）は

$$\mu_j(T) = {}_{n_1}C_{T_{0,0}} \times {}_{n_2}C_{T_{1,0}} \tag{1}$$

λ の確率は同じであるが、一つの偶現表の確率は同じではない。偶現表の確率は

$$Pr(T) = \frac{{}_{n_1}C_{T_{0,0}} \times {}_{n_2}C_{T_{1,0}}}{{}_nC_\xi} = \frac{n_1!\ n_2!\ \xi!\ (n-\xi)!}{n!\ T_{0,0}!\ T_{0,1}!\ T_{1,0}!\ T_{1,1}!}$$

である。周辺度数が固定されているとき、独立に動くランダム変数は一つであり（例えば、$T_{0,0}$）、これは Fisher の正確確率の計算の場合と同じ、超幾何分布に従う。

以上をまとめると、一つのマーカー j について、一つの偶現表 T は $\mu_j(T)$ 個の $\mathbf{d}_i$ に対応し、一つの $\mathbf{d}_i$ は $\xi!(n-\xi)!$ 個の ω に対応する。一つの ω の確率は $1/n!$ なので、偶現表 T の確率は

$$Pr(T) = {}_{n_1}C_{T_{0,0}} \times {}_{n_2}C_{T_{1,0}} \times \xi!(n-\xi)! \times \frac{1}{n!} = \frac{{}_{n_1}C_{T_{0,0}} \times {}_{n_2}C_{T_{1,0}}}{{}_nC_\xi}$$

である。

1.3 Pearson score、その最大スコアと求める確率

一つの λ（従って、一つの $\mathbf{d}_i$）が決まれば、任意の j について一つの偶現表が決まるので、それにより Pearson のスコア $S_j(\mathbf{d}_i)$ が決まる。一つの $\mathbf{d}_i$ について、すべての j の中で最大のスコアを

$$S(\mathbf{d}_i) = \max_j S_j(\mathbf{d}_i)$$

とする。

我々の目的は、観察された $S(\mathbf{d})$ 以上の $S(\mathbf{d}_i)$ を示す ω の集合の確率を求めることであるが、同じ λ の中の ω によってできる偶現表は皆同じなので、それにより計算されるスコアは同じであり、しかも λ は一つの $\mathbf{d}_i$ と対応し、その確率はすべて同じなので

$$Pr(\{\omega|S(\mathbf{d}_i) \geq S(\mathbf{d})\}) = Pr(\{\lambda|S(\mathbf{d}_i) \geq S(\mathbf{d})\}) = Pr(S(\mathbf{d}_i) \geq S(\mathbf{d}))$$

が、求める確率である。

すべての λ の（従って $\mathbf{d}_i$ の）集合を $\mathcal{F} = \{\mathbf{d}_i\}$ とする。

$\mathbf{d}_i$ のうち、$S(\mathbf{d}_i) \geq S(\mathbf{d})$ を満たすものの集合を $\mathcal{H}$ とする、即ち

$$\mathcal{H} = \{\mathbf{d}_i \in \mathcal{F}|\ S(\mathbf{d}_i) \geq S(\mathbf{d})\}$$

集合を構成する要素の数で表すと、我々の目的は

$$p = \frac{|\mathcal{H}|}{|\mathcal{F}|} = \frac{|\mathcal{H}|}{{}_nC_\xi} \tag{2}$$

を求めることである。従って、$|\mathcal{H}|$ を求めればよい。ただし、有限集合 A に関して、$|A|$ は要素の数を示す。ここで、$\mathcal{F}$, $\mathcal{H}$ を構成する要素 $\mathbf{d}_i$ の確率はすべて等しいことに注意。

1.4 Permutation event ごとの有意 SNP の数

$\mathcal{H}$ を構成する要素 $\mathbf{d}_i \in \mathcal{H}$ についてはすべて $S(\mathbf{d}_i) \geq S(\mathbf{d})$ を満たすので、最小 1 個の有意の SNP があることは確実である（j を一つの SNP と表現した）。そのような $\mathbf{d}_i$ について、$S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ を満たす j の数を

$$Q(\mathbf{d}_i) = |\{j|1 \leq j \leq m, S_j(\mathbf{d}_i) \geq S(\mathbf{d})\}|$$

とする。

$S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ を満たす j の数により次のように $\mathbf{d}_i$ に重み付けをする。

$$g(\mathbf{d}_i) = \frac{Q(\mathbf{d}_i)}{\sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i)} \tag{3}$$

また、j についての Pearson score $S_j(\mathbf{d}_i)$ により、次のような $\mathbf{d}_i$ の集合を定義する。

$\mathcal{H}_j = \{\mathbf{d}_i | S_j(\mathbf{d}_i) \geq S(\mathbf{d})\}$

$\mathbf{d}_i$ が $\mathcal{H}_j$ の要素であれば $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ であり、従って $S(\mathbf{d}_i) \geq S(\mathbf{d})$ なので必ず $\mathcal{H}$ の要素である。従って、$\mathcal{H}_j$ は $\mathcal{H}$ の部分集合。

$\mathcal{H}$ の要素の $\mathbf{d}_i$ について、いずれかの j で必ず $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ となる j があるので、その j について、必ず $\mathcal{H}_j$ の要素である。

従って、$\mathcal{H} = \cup_{j=1}^{m} \mathcal{H}_j$

1.5 $\mathcal{G}$-sampler

$\mathbf{d}_i \in \mathcal{H}$ を対象とし、次のような $\mathcal{G}$-sampler を構成する。

ここで、式（1）の関数 μ_j を次のように定義しなおす。ただし、S は本来 $\mathbf{d}_i$ の関数であるが、一つの偶現表 T に対応するすべての $\mathbf{d}_i$ について $S(\mathbf{d}_i)$ は同じなので、S を T の関数とみなすことができ、

$$\mu_j(T) = \begin{cases} \mu_j(T) & S(T) \geq S(\mathbf{d}) \\ 0 & otherwise \end{cases} \tag{4}$$

即ち、$\mathbf{d}_i \in \mathcal{H}$ と対応する T 以外には 0 を与える。

特定の座位 j における偶現表について、$C_j = \{T | S(T) \geq S(\mathbf{d})\}$ とする。

1. マーカー j を、確率 $|\mathcal{H}_j| / \sum_{a=1}^{m} |\mathcal{H}_j|$ で抽出する。
2. 偶現表 T を C_j から、確率 $\mu_j(T)/|\mathcal{H}_j|$ で抽出する。
3. 偶現表 T から一つの permutation event、$\mathbf{d}_i \in \mathcal{H}$ を等確率で、即ち、確率 $1/\mu_j(T)$ で抽出する。

このような $\mathcal{G}$-sampler は $\mathcal{H}$ の要素である $\mathbf{d}_i$ の一つを選択するものであり、それを一つの実験とすると一つの確率空間が形成される。しかも $\mathbf{d}_i \notin \mathcal{H}$ である $\mathbf{d}_i$ については確率 0 と考えれば、標本空間は同じ Ω と考えることもできる。以前の確率空間では $Pr(\omega) = 1/n!$ であり、また $Pr(\mathbf{d}_i) = 1/{}_nC_\xi$ であった。即ち、すべての permutation event は等確率で選択された。

今回の $\mathcal{G}$-sampler による確率空間は標本空間は同じであるが、確率測度は以前とは異なったものである。即ち、

$$Pr(\mathbf{d}_i) = \begin{cases} g(\mathbf{d}_i) & \mathbf{d}_i \in \mathcal{H} \\ 0 & \mathbf{d}_i \notin \mathcal{H} \end{cases} \tag{5}$$

なぜなら

$\mathbf{d}_i \in \mathcal{H}$ について $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ となる一つの j（一つ以上ある）がステップ 1 で抽出される確率は $|\mathcal{H}_j| / \sum_{a=1}^{m} |\mathcal{H}_j|$、その上で、$\mathbf{d}_i$ が抽出される確率は

$$\mu_j(T)/|\mathcal{H}_j| \times 1/\mu_j(T)$$

であるが、$\mathbf{d}_i$ は $Q(\mathbf{d}_i)$ 個の $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ となる j を持っているので、結局 $\mathbf{d}_i$ が抽出される確率は、

$$Q(\mathbf{d}_i) \times \frac{|\mathcal{H}_j|}{\sum_{a=1}^{m} |\mathcal{H}_j|} \times \frac{\mu_j(T)}{|\mathcal{H}_j|} \times \frac{1}{\mu_j(T)} = \frac{Q(\mathbf{d}_i)}{\sum_{a=1}^{m} |\mathcal{H}_j|}$$

すべての $\mathbf{d}_i$ について、有意な SNP 数を合計すると（即ち、$S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ なる j を総合計すると）、それは $\sum_{a=1}^{m} |\mathcal{H}_j| = \sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i)$ である。従って、上式は式（3）より $g(\mathbf{d}_i)$ に等しい。

このように、$\mathcal{G}$-sampler ではそれぞれの $\mathbf{d}_i$ が抽出される確率は異なるが、これはもともとの実験で $\mathcal{F}$、あるいは $\mathcal{H}$ の中の $\mathbf{d}_i$ が等確率で抽出される事とは異なることに注意すべきである。即ち、元々の実験では $\mathbf{d}_i$ を等確率で抽出しているが、$\mathcal{G}$-sampler では $Q(\mathbf{d}_i)$ 個の $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ となる j を持っている $\mathbf{d}_i$ は、1 個しか持っていない $\mathbf{d}_i$ よりも $Q(\mathbf{d}_i)$ 倍抽出されやすい。即ち、$\mathcal{G}$-sampler では $\mathbf{d}_i \in \mathcal{H}$ を等確率で抽出しているのではなく、すべての $\mathbf{d}_i \in \mathcal{H}$ に含まれる $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ となる j を等確率で抽出しているのである（即ち、有意な SNP を等確率で選択している）。このような $(\mathbf{d}_i, j)$ の集合を Λ とすると $\sum_{(\mathbf{d}_i,j)\in\Lambda} 1 = \sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i)$ であり、

$$|\mathcal{H}| = \sum_{(\mathbf{d}_i,j)\in\Lambda} \frac{1}{Q(\mathbf{d}_i)} \tag{6}$$

ここはちょっとわかりにくいと思うので、きわめてくだらない例で説明する。

有る年、日本のプロ野球で年間で n 個の本塁打が打たれたという。すべての打球が集められ、それぞれの球に打者の本塁打数が書かれている。例えば、長島は 34 本なので、長島の打った打球には 34 の数字が書かれている。これらの n 個の打球のデータから本塁打を打った選手の人数 m がわかるだろうか。i 番目の球に書いてある数字を s_i とすると、ホームラン打者数は

$$m = \sum_{i=1}^{n} \frac{1}{s_i}$$

これで、式（6）は理解できたであろうか。

もし、（復元抽出で）無限に球を抽出し、球に書かれた数の逆数の平均を取ると、大数の法則により、

$$\frac{m}{n} = E(\frac{1}{v_i}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{v_i}$$

ただし、v_i は i 番目に抽出された球に書かれていた数字である。

以上の考察から、$\mathcal{G}$-sampler では

$$E[\frac{1}{Q(\mathbf{d}_i)}] = \frac{|\mathcal{H}|}{|\Lambda|} = \frac{|\mathcal{H}|}{\sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i)}$$

$$|\mathcal{H}| = \sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i) \times E[\frac{1}{Q(\mathbf{d}_i)}] = \sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i) \times \lim_{N_R\to\infty} \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{1}{Q(\mathbf{d}_i)} \tag{7}$$

である。

また、$\mathcal{G}$-sampler で N_R 個のサンプリングを行い、その標本平均を $E[*]$ に代入したものが $|\mathcal{H}|$ の推定値であり、

$$|\hat{\mathcal{H}}| = \sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i) \times \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{1}{Q(\mathbf{d}_i)}$$

$\mathcal{G}$-sampler の 3 つのステップの中で、(3) は容易である。特定の j の偶現表 T が選択されれば、マーカー j について genotype 1 の n_1 人、genotype 2 の n_2 人からそれぞれ $T_{0,0}$ 人、$T_{1,0}$ 人（表 1）をランダムに抽出すればよい。

しかし、ステップ 1,2 は結構むずかしい。それは 1. $\sum_{j=1}^{m} |\mathcal{H}_j|$, $|\mathcal{H}_j|$ を計算する必要があること、2. 偶現表 T を C_j から $\mu_j(T)/|\mathcal{H}_j|$ の確率で抽出する必要があることがネックになるからである。

1.5.1 正確法

まず、正確法による抽出法を紹介する。

一つの j について、我々はすべての可能な偶現表（O^{s-1}）を数え上げ、集合 C_j を構築することが出来る。また、それぞれの偶現表 T について式（1）により $\mu_j(T)$ を計算することが出来、さらに $|\mathcal{H}_j|$ を $|\mathcal{H}_j| = \sum_{T \in C_j} \mu_j(T)$ により計算することが出来る。この計算のオーダーは $O(n^{s-1}m + N_R nm)$ である。

1.5.2 近似アルゴリズム

$\sum_{j=1}^{m} |\mathcal{H}_j|$ を計算するためにまず、定数 β を決める。m 個のマーカーからランダムに β 個のマーカーを選択し、その集合を L とする。続いて L の中のマーカー j について上記の正確法により $|\mathcal{H}_j|$ を計算する。

引き続き、$m/\beta \sum_{j \in L} |\mathcal{H}_j|$ により $\sum_{j=1}^{m} |\mathcal{H}_j|$ を推定する。

続いて、C_j から T を $\mu_j(T)/\sum_{T \in C_j} \mu_j(T)$ の確率で抽出するためには Metropolis-Hastings アルゴリズムを用いる。

式（1）の確率で T が抽出される MCMC アルゴリズムを構成する。注意が必要なのは、可能な状態 T について、各セルが負ではないという条件の他に、$T \in C_j$ という条件が必要なことである。

1.5.3 MH アルゴリズムによる偶現表のサンプリング

表 1 は変数が面倒なので、次の偶現表について考える。

Status	genotype 1	genotype 2	total
disease	a	b	m_1
control	c	d	m_2
total	n_1	n_2	n

特定の座位 j について、周辺度数、$n_1,\ n_2,\ m_1,\ m_2$ は固定されており、独立変数は 1 つであり、a に相当するセルの値をランダム変数 X とすると、X は超幾何分布に従い、その確率は

$$Pr(X=a)=\frac{n_1!\ n_2!\ m_1!\ m_2!}{n!\ a!\ b!\ c!\ d!} \tag{8}$$

X の動く範囲は、特に制限が無い場合は

$\max(0,a-d)\le X\le \min(m_1,n_1)$ である。$x_l=\max(0,a-d),\ x_u=\min(m_1,n_1)$ とする。この範囲で動いたときの偶現表の集合を C_t とする。周辺度数が固定されている場合、偶現表は X の一対一対応するので、C_t、あるいは C_j は X の動ける範囲と同一視できる。

この X の範囲で、式（8）の確率で特定の偶現表 $T\in C_t$ が得られるマルコフ連鎖を作るには、次のように偶現表を選択し、ステップを進めると良い。C_t が状態空間であり、その要素である偶現表（あるいは X の値）が状態である。

1. $k=0$ とし、特定の偶現表 $T^{(k)}=T\in C_t$ を選択する。
2. 確率 $1/2$ で $T^{(k)}$ の a を一つ増やす $(i=1)$ か減らす $(i=-1)$ か選択する。
3. 周辺度数を変えないように $b,\ c,\ d$ も変化させ、新たに出来た偶現表を $T'^{(k+1)}$ とする。
4. $T'^{(k+1)}\in C_t$ を確かめる（$a,\ b,\ c,\ d$ が負にならないかを確かめる。$x_l\le a\le x_u$ を調べるのと同じ）。
5. $T'^{(k+1)}\notin C_t$ なら、同じ状態（$T^{(k+1)}=T^{(k)}$）を選択しステップを進め（k を 1 つ増やす）、(7) に跳ぶ。
6. $T'^{(k+1)}\in C_t$ なら次の値を計算する。

$$y=Pr(T'^{(k+1)})/Pr(T^{(k)})=\begin{cases}\frac{bc}{(a+1)(d+1)} & i=1\\ \frac{ad}{(b+1)(c+1)} & i=-1\end{cases} \tag{9}$$

   $\min[1,y]$ の確率で $T^{(k+1)}=T'^{(k+1)}$ とし、ステップを進め（k を 1 つ増やす）、(7) に跳び、$1-\min[1,y]$ の確率で、状態を同じとし（$T^{(k+1)}=T^{(k)}$）、ステップを進める（k を 1 つ増やす）。
7. (2) に戻る。

1.5.4 $T \in C_j$ の範囲で動くマルコフ連鎖

しかし、我々は $T \in C_t$ ではなく、$T \in C_j$ の範囲で動くマルコフ連鎖を考えなければならない。そして、T が次の確率で表れるようにすれば良い。

$$Pr(X = a) = \frac{n_1!\ n_2!\ m_1!\ m_2!}{n!\ a!\ b!\ c!\ d!\ Pr(C_j)} \tag{10}$$

C_j は Pearson score が $S(\mathbf{d})$ 以上の座位 j における偶現表の集合であるが、Pearson score は

$$\frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(d+b)(d+c)}$$

であり、周辺度数が固定されていた場合、$(bc-ad)^2$ のみが変化する部分である。a と d、b と c は増加、減少が同時に起きるので、Pearson score は $X = a$ が小さいか大きい場合に大きくなり、C_j に含まれる偶現表に対応する X は、$x_l \leq X \leq x_p$、$x_q \leq X \leq x_u$、あるいは $x_l \leq X \leq x_p$ or $x_q \leq X \leq x_u$ の範囲になる。前の 2 つの例では X が動ける範囲が一つの範囲にまとまっているので比較的容易であるが、最後の例では X が二つの範囲に分離しているので、その二つの範囲の間を跳びわたるために工夫が必要である。

1.5.5 X の動ける範囲が二つに分離していない場合

まず、X の動ける範囲が二つに分離していない場合は、次のようなマルコフ連鎖を作ればよい。

1. $k = 0$ とし、特定の偶現表 $T^{(k)} = T \in C_j$ を選択する。
2. 確率 $1/2$ で $T^{(k)}$ の a を一つ増やす $(i = 1)$ か減らす $(i = -1)$ か選択する。
3. 周辺度数を変えないように $b,\ c,\ d$ も変化させ、新たに出来た偶現表を $T'^{(k+1)}$ とする。
4. $T'^{(k+1)} \in C_j$ を確かめる。
5. $T'^{(k+1)} \notin C_t$ なら、同じ状態（$T^{(k+1)} = T^{(k)}$）を選択しステップを進め（k を 1 つ増やす）、(7) に跳ぶ。

6. $T'^{(k+1)} \in C_t$ なら次の値を計算する。

$$y = Pr(T'^{(k+1)})/Pr(T^{(k)}) = \begin{cases} \frac{bc}{(a+1)(d+1)} & i = 1 \\ \frac{ad}{(b+1)(c+1)} & i = -1 \end{cases} \tag{11}$$

$\min[1, y]$ の確率で $T^{(k+1)} = T'^{(k+1)}$ とし、ステップを進め（k を 1 つ増やす）、(7) に跳ぶ。$1 - \min[1, y]$ の確率で、状態を同じとし（$T^{(k+1)} = T^{(k)}$）、ステップを進める（k を 1 つ増やす）。

7. (2) に戻る。

1.5.6 X の動ける範囲が二つに分離している場合

X の動ける範囲が二つに分離している場合は、やや複雑である。X の動ける範囲が $x_l \leq X \leq x_p$ or $x_q \leq X \leq x_u$ のように分離する場合、次のようなマルコフ連鎖を作ればよい。

1. $k = 0$ とし、特定の偶現表 $T^{(k)} = T \in C_j$、即ち、上の範囲で一つの X の値を選択する。

2. 確率 $1/2$ で $T^{(k)}$ の a を一つ増やす $(i = 1)$ か減らす $(i = -1)$ か選択し、新たな a を a' とする。周辺度数を変えないように $b,\ c,\ d$ も変化させ、新たに出来た偶現表を $T'^{(k+1)}$ とする。

3. $T'^{(k+1)} \in C_j$ を確かめる。即ち、$T'^{(k+1)}$ の $X = a'$ が $x_l \leq a' \leq x_p$ or $x_q \leq a' \leq x_u$ を満たすかを確かめる。

4. $a' < x_l$ or $a' > x_u$ なら、同じ状態（$T^{(k+1)} = T^{(k)}$）を選択しステップを進め（k を 1 つ増やす）、(7) に跳ぶ。

5. $x_l \leq a' \leq x_p$ or $x_q \leq a' \leq x_u$ なら次の値を計算する。a, b, c, d は更新前の値。

$$y = Pr(T'^{(k+1)})/Pr(T^{(k)}) = \begin{cases} \frac{bc}{(a+1)(d+1)} & i = 1 \\ \frac{ad}{(b+1)(c+1)} & i = -1 \end{cases} \tag{12}$$

$\min[1, y]$ の確率で $T^{(k+1)} = T'^{(k+1)}$ とし、ステップを進め（k を 1 つ増やす）、(7) に跳ぶ。$1 - \min[1, y]$ の確率で、状態を同じとし（$T^{(k+1)} = T^{(k)}$）、ステップを進め（k を 1 つ増やす）、(7) に跳ぶ。

6. $x_p < a' < x_q$ なら区間を跳躍する必要が生じる。次の値を計算する。

$$r = Pr(X = x_q)/Pr(X = x_p) = \frac{a_p!\ b_p!\ c_p!\ d_p!}{a_q!\ b_q!\ c_q!\ d_q!} \tag{13}$$

ただし、$a_p,\ b_p,\ c_p,\ d_p$ は $a = x_p$ の時の $a,\ b,\ c,\ d$ の値、$a_q,\ b_q,\ c_q,\ d_q$ は $a = x_q$ の時の $a,\ b,\ c,\ d$ の値である。この式は、分母と分子を適当に約分することにより計算を簡略化できる。また、この値は一つのマルコフ連鎖で一回だけ計算すればよい。

区間を跳躍した後の状態を新たに定める必要があり、

$$a' = \begin{cases} x_q & i = 1 \\ x_p & i = -1 \end{cases} \tag{14}$$

とし、b, c, d もそれに従い更新した偶現表を新たに、$T'^{(k+1)}$ とする。

$$y = Pr(T'^{(k+1)})/Pr(T^{(k)}) = \begin{cases} r & i = 1 \\ 1/r & i = -1 \end{cases} \tag{15}$$

を計算し、$\min[1, y]$ の確率で $T^{(k+1)} = T'^{(k+1)}$ とし、ステップを進め（k を 1 つ増やす）、(7) に跳ぶ。$1 - \min[1, y]$ の確率で、状態を同じとし（$T^{(k+1)} = T^{(k)}$）、ステップを進め（k を 1 つ増やす）、(7) に跳ぶ。ただしあらかじめ計算した r について、$r \geq 1$ であれば $x_p \to x_q$、$r \leq 1$ であれば $x_q \to x_p$ への跳躍は無条件で行うことができる。

7. (2) に戻る。

正確法、近似法のいずれの場合にも、j について T が選択された後、$\mathbf{d}_i$ を選択し、$\mathbf{d}_i$ が抽出された後に $Q(\mathbf{d}_i)$ を計算する。n が小さな場合にはすべての偶現表の確率を計算することも可能である。

1.6 $g(\mathbf{d}_i)$ とその P 値

$g(\mathbf{d}_i)$ の計算のためには $\mathcal{G}$-sampler により選択された $\mathbf{d}_i$ について $Q(\mathbf{d}_i)$ を計算する必要がある。$\Gamma = \sum_{\mathbf{d}_i \in \mathcal{H}} Q(\mathbf{d}_i) = \sum_{j=1}^{m} |\mathcal{H}_j|$ は一度だけ計算すればよい。

まず

$$\Phi = \frac{1}{f(\mathbf{d}_i)} = {}_nC_\xi$$

を計算する。

式（2, **??**）より、$\mathcal{G}$-samper により選択された $\mathbf{d}_i$ について

$$p = \frac{|\mathcal{H}|}{|\mathcal{F}|} = \frac{\Gamma}{\Phi} \lim_{N_R \to \infty} \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{1}{Q(\mathbf{d}_i)}.$$

この p の推定の正確さは $1/Q(\mathbf{d}_i)$ の分散によっている。

1.7 LD decay を用いた計算の簡略化

上記の例では j を選択した後、偶現表 T を選択し、続いて $\mathbf{d}_i$ を選択した後、$Q(\mathbf{d}_i)$ を調べなければならず、そのためには、すべての k 番座位について $S_k(\mathbf{d}_i) \geq S(\mathbf{d})$ を調べなければならない。しかし、例えば 50 万の座位についてこれを調べるには時間がかかる。実は、特定の j について $S_j(\mathbf{d}_i) \geq S(\mathbf{d})$ の時、j から遠く離れた、あるいは別の染色体上の座位 k については $S_k(\mathbf{d}_i) \geq S(\mathbf{d})$ となる出来事は j 座位における出来事とは独立である。一般に、染色体上で、SNP 数 c 個以上離れた座位が独立（連鎖不平衡なし）とし、これを連鎖上限とする。そうなれば j から c SNP 数以内の範囲の $2c$ 個の座位 k についてのみ $S_k(\mathbf{d}_i) \geq S(\mathbf{d})$ をテストし、それ以外の $m - 2c - 1$ 個の座位（m は全 SNP 数）については座位あたり q の確率で $S_k(\mathbf{d}_i) \geq S(\mathbf{d})$ となると考えれば良い。q は一回だけ、最初に調べればよい。そして、j 番座位から c SNP 数を超えて離れている $m - 2c - 1$ 座位について、有意となる座位数の期待値は $(m - 2c - 1)q$ と考える。従って、$2c$ 個の座位について実際に有意性を計算した後、期待値 $(m - 2c - 1)q$ の Poisson 分布でサンプリングを行い、$2c$ 個の座位で有意であった座位数に加えればよい。

c を増やせば精度はあがるが、計算時間が長くなる。

参考文献

[1] Ito T, Inoue E, Kamatani N. Association test algorithm between a qualitative phenotype and a haplotype or haplotype set using simultaneous estimation of haplotype frequencies, diplotype configurations and diplotype-based penetrances. Genetics. 2004 168:2339-48.

[2] Furihata S, Ito T, Kamatani N. Test of association between haplotypes and phenotypes in case-control studies: examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. Genetics. 2006 174:1505-16.

[3] Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N. Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 2004 168:525-39.

分散因子分析による量的形質座位の同胞解析

第 12 章 ― 12.4 同胞対を用いた QTL 解析：ヘイスマン-エルストン法 ―

1　分散因子分析による量的形質座位の同胞解析（第 12 章、12.4）

2　QTL 解析の分散因子分析法

2.1　隠れマルコフモデルで各点の π_{ij}, δ_{ij} を求める場合

主として Pratt による方法を述べる [1]。まず、一つの家系について考える。今回は、QTL 表現型 X を次のモデルで表すとし: $X = g + G + \sum_i \beta_i K_i + e$、ここで g は今テストする座位と連鎖している主要遺伝子による効果を示すランダム変数とし G は連鎖していない座位による効果を表すランダム変数とする。e はその他の非遺伝的要因による効果を表すランダム変数とする。β_i は測定された変数 K_i の回帰係数や母平均などを含む定数とする（例えば、男女を示す変数 1,0 などを考えると良い。男女の差を表す値が β_i となる。母平均の項では変数は常に 1、β_i は母平均とすれば良い）。

これまでの議論では g, e のみを問題としたことになる。即ち g は Ge、e は Er に相当した。ここで説明変数を加える意味を考えてみる。

表 2　QTL の分散因子分析法の各説明変数と各要素との関係

説明変数	主要座位の遺伝子型	家族関係	個体
g	関係あり	関係あり	関係あり
G	関係なし	関係あり	関係あり
$\beta_i K_i$	関係なし	関係なし	関係あり
e	関係なし	関係なし	関係なし

上の表を説明する。G が g の関与する QTL 座位とは別の染色体の（連鎖していない）座位による効果とすると、個体において二つの座位の遺伝子型は独立である（メンデルの独立の法則）。しかし、同じ家系内の異なった個体について、G の関係する遺伝子型は独立ではない。今、その座位のマーカー遺伝子型のデータを取り扱っていない場合は、同じ家系内の異なった個体について G の関係する遺伝子型の関係（従って遺伝子型値の共分散）は家族関係のみによる。$\beta_i K_i$ の項は家族関係にも g の関係する QTL 遺伝子座の遺伝子型にも関係せず、各個体に固有のものである。e は個体にも関係しないランダム変数である。

これらのランダム変数は平均はいずれも 0、分散は $\sigma_g^2, \sigma_G^2, \sigma_e^2$ の正規分布に従うとする。即ち、g, G は前述の遺伝子型値変位に相当するように変換されたものとする。遺伝的分散 σ_g^2, σ_G^2 はいずれも相加的分散とドミナンス（非相加的）分散に分解できるとする。即ち、$\sigma_g^2 = \sigma_{ga}^2 + \sigma_{gd}^2$、$\sigma_G^2 = \sigma_{Ga}^2 + \sigma_{Gd}^2$。もし、$g, G, e$ が無相関とすると、X の分散は、$\sigma_{ga}^2 + \sigma_{gd}^2 + \sigma_{Ga}^2 + \sigma_{Gd}^2 + \sigma_e^2$。

今、同一家系の二人の個体 i, j の表現型（X_i, X_j) の共分散を考える。g, G, e は無相関と仮定したので、

$$Cov(Xi, Xj) = \begin{cases} \sigma_{ga}^2 + \sigma_{gd}^2 + \sigma_{Ga}^2 + \sigma_{Gd}^2 + \sigma_e^2 & \text{if } i = j \\ \pi_{ij}\sigma_{ga}^2 + \delta_{ij}\sigma_{gd}^2 + 2\Phi_{ij}\sigma_{Ga}^2 + \Delta_{ij}\sigma_{Gd}^2 & \text{if } i \neq j \end{cases},$$

ここで、π_{ij} は個体 i, j 間の主要座位における ibd であるアレルの割合（観察データのもとでの)、δ_{ij} はその座位での両方のアレルが ibd である確率（観察データのもとでの）とする。主要座位の遺伝子型の効果（遺伝子型値）の共分散が $\pi_{ij}\sigma_{ga}^2 + \delta_{ij}\sigma_{gd}^2$ となることについては前述の式**??**に示した。

また、Φ_{ij} は個体 i, j の親縁係数（遺伝子型データの無い家系内の二人の個体のそれぞれからアレルを一つずつ取ったとき、それが ibd である確率。これは家系図からの ibd であるアレルの割合、即ち関係係数の半分である)、Δ_{ij} は二人が両方のアレルとも ibd である確率（家系図からのみわかる)、即ち近交係数である。主要な座位と連鎖していない QTL 座位については、π_{ij} の期待値は $2\Phi_{ij}$、δ_{ij} の期待値は Δ_{ij} なので、それらの効果による分散因子は $2\Phi_{ij}\sigma_{Ga}^2 + \Delta_{ij}\sigma_{Gd}^2$ となることについては前述の式**??**に示した。

家系図がわかれば Φ_{ij}, Δ_{ij} はわかる。連鎖しているマーカー座位の遺伝子型の情報と家系図から隠れマルコフモデルを用いた方法で（genehunter にインストールされている方法)、すべてのマーカー座位の情報を用いた上での各点の π_{ij}, δ_{ij} はわかる。そこで、ゲノム上のある点について、各分散が（変数として）与えられれば、各々の家系について QTL 表現型の分散共分散行列 $\mathbf{V}$ は作れる。

r 番目の家系の構成員の QTL 表現型を要素とするベクトル（表現型ベクトル）$\mathbf{X}_r$ を考える。個々の構成員の QTL 表現型の結合分布について多変量正規分布が仮定できるとすると、全家系のデータの尤度は

$$\log L = c - \frac{1}{2}\sum_{r=1}^{R} \log[\det(\mathbf{V}_r)] - \frac{1}{2}\sum_{r=1}^{R} (\mathbf{X}_r - \mathbf{K}_r\boldsymbol{\beta})' \mathbf{V}_r^{-1} (\mathbf{X}_r - \mathbf{K}_r\boldsymbol{\beta}) \tag{1}$$

ここで、c は定数、det は行列式、$\mathbf{V}_r$ は r 番目の家系の分散共分散行列（行列の要素は分散因子を変数として含む)、$\mathbf{K}_r$ は r 番目の家系の変数行列（これは男性、女性などにより、個体に固有な値

として定まるものである)、R は分析する家系数、$\boldsymbol{\beta}$ は固定効果（回帰係数など）のベクトル（これは未知）である。

この尤度を最大化するパラメータ（各分散因子と $\boldsymbol{\beta}$）の推定は Fisher のスコア法[1]により行われる。スコア法については、文献 [2] を参照。

意味の無い推定を避けるため、分散因子はすべて非負の拘束条件を設定する。

π_{ij}, δ_{ij} は、すべてのマーカー座位のデータを用いて隠れマルコフモデルで計算できる各点における継承ベクトルの事後分布から計算できる。この場合、個体 i, j 間の ibd が 0,1,2 である確率 z_0, z_1, z_2 とは次の関係にある。

$\pi_{ij} = 0.5z_1 + z_2, \delta_{ij} = z_2$

次に、任意のある点に QTL 座位があるという仮説を検定する。即ち特定の場所への連鎖の検定である。その場所での最尤値と σ^2_{ga} と σ^2_{gd} を 0 と置いた時（即ち連鎖無し）の最尤値との比を用いる。

最も単純なモデルでは σ^2_{ga} のみをモデル化する。$\sigma^2_{Ga}, \sigma^2_{Gd}, \sigma^2_{gd}$ は最初から 0 とする。この場合 $\log_e$-尤度比の 2 倍の分布は漸近的に χ^2_1 変数と 0 での point mass の $\frac{1}{2}$:$\frac{1}{2}$ の混合である [3]。一変数以上の変数をテストする場合の尤度比の分布は良くわかっていないが、一般的にやはり χ^2 変数の混合である。相加的分散とドミナンス分散の両方を未知数とする場合、保守的なアプローチを用い、テスト統計量を χ^2_2 分布と比較する　。

実際には、分散因子のうち、σ_{ga}, σ_{gd}(これはしばしば最初から 0 とされる), σ_{Ga}, σ_e のみで、$\boldsymbol{\beta}$ は母平均のみを未知のパラメータとして推定する。

2.2　QTL 解析の分散因子分析法: 隠れマルコフモデルを用いない場合

これまでに述べた方法では QTL 座位とマーカー座位との組換え割合 θ は $\mathbf{V}$ の中には入ってこない。隠れマルコフモデルで π_{ij}, δ_{ij} を計算する過程ですべての遺伝子型データの情報は考慮されており、QTL 座位の存在しうるすべての点上での $\mathbf{V}$ を計算できるからである。

マーカー座位上での、その座位の遺伝子型データのみを用いて π_{ij}, δ_{ij} を計算した場合には QTL 座位とマーカー座位の間の組換え割合を考慮する必要がある [4]。この場合、分散因子は QTL 座位に

[1] スコア法は、Newton-Raphson 法の 2 階導関数の行列を、その期待値の行列で置き換える方法である。種々のパッケージが公表されている。

マーカー座位がある場合とは変わってくる。例えば、X_i, X_j の共分散は相加的部分については、QTL座位がマーカー座位の上にある場合は前述のように $\pi_{ij}\sigma_b^2$ であるが、i, j が sib pair で QTL 座位とマーカー座位の間の組換え割合が θ の場合、共分散の相加的部分は $[1/2 + (1-2\theta)^2(\pi - 1/2)]\sigma_b^2$ である。従って、$\mathbf{V}_r$ の要素に各分散因子の他に θ が入ることになる。推定はやはり前述の数式 1 を最大化する。分散因子が 0 であるという帰無仮説をテストするには、拘束の無い条件:例えば $\sigma_b^2, \sigma_G^2, \sigma_e^2$、場合によっては θ を推定する条件での最尤値と帰無仮説（$\sigma_b^2 = 0$）での最尤値の比の自然対数の-2倍が漸近的に χ^2 分布に従うことを用いる。自由度は拘束したパラメータの数である。

文献 [5] に一般の家系においてペアの表現型の差の平方の変わりに、共分散を用いて ibd への回帰分析を行う方法が発表されている。

参考文献

[1] Pratt SC, Daly MJ, Kruglyak L (2000) Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. Am J Hum Genet 66: 1153-1157

[2] Lange K, Westlake J, Spence MA (1976) Extensions to pedigree analysis: III. Variance components by the scoring method. Ann Hum Genet 39, 485-491

[3] Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc 82:605-610

[4] Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535-543

[5] Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19: 1-17

二段階絞込みによる SNP 探索の有意水準と検出力

第 13 章 － 13.3 多数の SNP を用いた関連解析 －

1　二段階絞込みによるSNP探索の有意水準と検出力（第13章、13.3）

まず、二つの段階にわけてSNPを絞り込む場合、個々の段階の有意水準と全体を通した最終的なタイプIの過誤の確率の関係を検討する。第一段階の有意水準をα_1とする。即ち、第一段階の検定でP値がα_1以下であった場合のみ、第二段階でそのSNPをタイピングし、データを用いて検定する。第二段階ではいくつか異なった方法によりP値を計算し、下記に述べる方法でα_2(独立法)、γ（P値積法)、α_u（joint法）を用い有意性を検討する。このような一連の解析過程で、結果的に最終的なタイプIの過誤の確率がどうなるかを検討する。即ち、一つのあるSNPが最終的に有意であるとして結論された場合、帰無仮説の下で（関連が無いという仮定で）それが有意と判断される確率を計算するのである。次に、第二段階での有意水準を調製することにより、最終的なタイプIの過誤の確率が求める適当な値になるようにし、その有意水準の下で検出力の推定を行う。検出力の推定のためには、$p_1 \neq p_2$の条件を与えて有意になる確率を計算する。この時のオッズ比は$[p_1/(1-p_1)]/[p_2/(1-p_2)]$である。

二段階絞込みを行う場合、以下のような方法が考えられる。

1.1　二つの段階の検定を独立として処理（独立法）

独立法は、第一段階の検定と第二段階の検定を全く独立のものと考える方法である。Replication methodと呼ぶ場合も有る。第一段階の有意水準をα_1とし、有意のSNPのみを第二段階でタイピングを行う。第二段階のデータの検定で有意水準α_2で有意であれば最終的に有意のSNPとする。

この方法（独立法）で検定を行った場合、実際のタイプIの過誤の確率は$\alpha_1\alpha_2$である。即ち、独立法の場合、最終的なタイプIの過誤の確率は第一段階、第二段階の有意水準のみによって決まる。これを一つのSNPに対する最終的な有意水準とすればよい。従って、第一段階での有意水準をα_1と決め、最終的な有意水準をPとしたければ、第二段階での有意水準を$\alpha_2 = P/\alpha_1$とすればよい。

ただし、第一段階と第二段階で特定の遺伝子型（例えば、リスク型）の割合が逆転している場合は研究者は有意のSNPとは考えない事を考慮すると、実際のタイプIの過誤の確率は$\alpha_1\alpha_2/2$となる。

例えば、症例と対照におけるリスク型の遺伝子型頻度の違いを検定する場合、二群間の母比率の違いの検定を行うことが多い（第**??**節)。この場合用いられる統計量は以下のようなものである（式

??)。

$$F = \frac{X}{n_1} - \frac{Y}{n_2} \sim N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}) \tag{1}$$

この式の F の符号が第一段階と第二段階で異なった場合、研究者は意味の無い SNP と判断するであろう。帰無仮説の下で F の符号が異なる確率は $1/2$ なので、実際の全体のタイプ I の過誤の確率は $\alpha_1\alpha_2/2$ と考えられるのである。従って、この場合、全体のタイプ I の過誤の確率を P にしたければ、$\alpha_2 = 2P/\alpha_1$ とすればよい。

検出力を計算するためには対立仮説の下で有意となる確率を計算する。

Case と control の二つの標本を選択する場合、特定の SNP 情報を持つ個体の割合が case と control で母集団において $p_1,\ p_2$ と異なる場合の検出力の計算法（二群間の母比率の違いの検定）については前述した（第**??**節)。即ち、まず最終的なタイプ I 過誤率が適当な値となるように有意水準 α を定める。次に、二つの母集団における特定の SNP 情報を保有する個体の割合 $p_1,\ p_2,\ (p_1 > p_2)$ を与え、式（**??**）を用いて検出力を計算する。ここで求められた検出力 G が、その有意水準で有意となる確率である。従って、第一段階の有意水準 α_1 の下での検出力を G_1、第二段階の有意水準 α_2 の下での検出力を G_2 とすると、全体の検出力は G_1G_2 である。

式（**??**）より、有意水準 α とした時の症例、対照研究の検出力は

$$\begin{aligned} G \quad &\simeq 1 - \Phi\{[\sqrt{p_c(1-p_c)(\frac{1}{n_1} + \frac{1}{n_2})}\ \Phi^{-1}(1-\alpha/2) - p_1 + p_2] \\ &\quad / \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\} \end{aligned} \tag{2}$$

なので、独立法の検出力は

$$\begin{aligned} W_i \quad &= \quad \simeq (1 - \Phi\{[\sqrt{p_c(1-p_c)(\frac{1}{n_1} + \frac{1}{n_2})}\ \Phi^{-1}(1-\alpha_1/2) - p_1 + p_2] \\ &\qquad / \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\}) \\ &\quad \times \quad (1 - \Phi\{[\sqrt{p'_c(1-p'_c)(\frac{1}{n'_1} + \frac{1}{n'_2})}\ \Phi^{-1}(1-\alpha_2/2) - p_1 + p_2] \\ &\qquad / \sqrt{\frac{p_1(1-p_1)}{n'_1} + \frac{p_2(1-p_2)}{n'_2}}\}) \end{aligned} \tag{3}$$

ただし、$p_c,\ p'_c$ はそれぞれ第一段階、第二段階での推定された帰無仮説の下でのリスク型遺伝子型の

頻度、n_1', n_2' は第二段階での症例、対照のサンプルサイズである。p_c, p_c' は、式（**??**）により計算される。

1.2 二つの P 値の積を有意水準と比較（P 値積法: Multiplied P value method）

上述の方法では第一段階での P 値の情報は α_1 より低いという部分のみが用いられている。第一段階での P 値の情報を最終段階での判断にも用いれば検出力が上昇する可能性がある。

次の方法は、以下のようである。第一段階で計算した P 値、P_1 が α_1 の有意水準で有意であったときのみ第二段階でタイピングを行う。第二段階のデータのみで計算した P 値である P_2 を用い、最終的な有意性は $P_1P_2 \leq \gamma$ により判定する。

この方法について、最終的なタイプ I の過誤の確率は γ で良いように思えるかも知れないが、実はそうではない。第一段階で P 値が α_1 よりかなり低いと、第二段階では P 値が大きくても最終的に有意と判断される。実際の最終的なタイプ I の過誤の確率は γ より大きくなる。ここで数理的に帰無仮説の下でのタイプ I の過誤の確率を計算しよう。

第一段階、第二段階とも前述の二集団間の母比率の差の検定（第**??**節）を行うこととする。式（**??**）より、帰無仮説が正しいとすると（二つの集団の母比率が等しいとき）第一段階、第二段階の統計量、

$$\begin{aligned} Z_1 &= (\frac{X_m}{m_1} - \frac{Y_m}{m_2}) / \sqrt{p_c(1-p_c)(\frac{1}{m_1} + \frac{1}{m_2})} \sim N(0,1) \\ Z_2 &= (\frac{X_n}{n_1} - \frac{Y_n}{n_2}) / \sqrt{p_c'(1-p_c')(\frac{1}{n_1} + \frac{1}{n_2})} \sim N(0,1) \end{aligned} \tag{4}$$

は独立で、標準正規分布に従う。ただし、X_m, Y_m は第一段階で、症例、および対照において、特定の（例えばリスク型）の遺伝子型を持つ個体数、X_n, Y_n は第二段階で、症例、および対照において、特定の（例えばリスク型）の遺伝子型を持つ個体数である。p_c, p_c' はそれぞれ第一段階、第二段階のサンプルから推定された帰無仮説における比率（例えばリスク型の遺伝子型の頻度）であり、m_1, m_2 は第一段階での第一、第二集団のサンプルサイズ、n_1, n_2 は第二段階での第一、第二集団のサンプルサイズである。

P 値積法で最終的に有意と判断される SNP の条件は、$|z_1| \geq 1 - \alpha/2$ の条件の下で、

$$\Phi(-|z_1|)\Phi(-|z_2|) = [1 - \Phi(|z_1|)][1 - \Phi(|z_2|)] \leq \frac{\gamma}{4} \tag{5}$$

である。

式（5）は $z_1,\ z_2$ について、0 を軸として対象なので、$z_1 \geq 0,\ z_2 \geq 0$ のときのみの確率を考え、上式の条件を満たす確率を計算したい場合は、4 倍すればよい。

1. 式（5）を満たす $Z_1,\ Z_2$ の確率は $z_1 \geq 0,\ z_2 \geq 0$ のとき

   $z_1 \geq \Phi^{-1}(1-\alpha_1/2)$ の条件の下で、

$$[1-\Phi(z_1)][1-\Phi(z_2)] \leq \frac{\gamma}{4} \tag{6}$$

   が満たされる確率である。変形し

$$1-\Phi(z_2) \leq \frac{\gamma}{4[1-\Phi(z_1)]}$$

   さらに、

$$\Phi(z_2) \geq 1-\frac{\gamma}{4[1-\Phi(z_1)]} \tag{7}$$

   であるが、$z_2 \geq 0$ なので、$\Phi(z_2) \geq 1/2$ であり、

$$1-\frac{\gamma}{4[1-\Phi(z_1)]} \leq \frac{1}{2}$$

   の場合、つまり $\Phi(z_1) \geq 1-\gamma/2$ の場合、即ち、$z_1 \geq \Phi^{-1}(1-\gamma/2)$ の場合、式（6）は常に満たされる。

   それ以外、即ち、$0 \leq z_1 \leq \Phi^{-1}(1-\gamma/2)$ の場合は、式（7）より、

$$z_2 \geq \Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\}$$

   が、式（6）を満たす条件である。ゆえに、求める確率は

$$P_{global++} = \frac{1}{2}\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\phi(z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\phi(z_1)\frac{\gamma}{4[1-\Phi(z_1)]}dz_1 \tag{8}$$

   であるが、

$$\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\phi(z_1)dz_1 = \gamma/2$$

   および不定積分、

$$\int \phi(z_1)\frac{\gamma}{4[1-\Phi(z_1)]}dz_1 = -\frac{\gamma}{4}\log[1-\Phi(z_1)] + C$$

より、

$$P_{global++} = \frac{\gamma}{4} + \frac{\gamma}{4}\log\frac{\alpha_1}{\gamma}$$

これは、求める全体の確率の 1/4 なので、

$$P_{global} = \gamma(1 + \log\frac{\alpha_1}{\gamma}) \tag{9}$$

2. 式（5）を満たす $Z_1,\ Z_2$ の確率は $z_1 \geq 0,\ z_2 \leq 0$ のとき

$z_1 \geq \Phi^{-1}(1-\alpha_1/2)$ の条件の下で、

$$[1-\Phi(z_1)]\Phi(z_2) \leq \frac{\gamma}{4} \tag{10}$$

が満たされる確率である。変形し

$$\Phi(z_2) \leq \frac{\gamma}{4[1-\Phi(z_1)]} \tag{11}$$

であるが、$z_2 \leq 0$ なので、$\Phi(z_2) \leq 1/2$ であり、

$$\frac{\gamma}{4[1-\Phi(z_1)]} \geq \frac{1}{2}$$

の場合、つまり $\Phi(z_1) \geq 1-\gamma/2$ の場合、即ち、$z_1 \geq \Phi^{-1}(1-\gamma/2)$ の場合、式（10）は常に満たされる。

それ以外、即ち、$0 \leq z_1 \leq \Phi^{-1}(1-\gamma/2)$ の場合は、式（11）より、

$$z_2 \leq \Phi^{-1}(\frac{\gamma}{4[1-\Phi(z_1)]})$$

が、式（14）を満たす条件である。ゆえに、求める確率は

$$P_{global+-} = \frac{1}{2}\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\phi(z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\phi(z_1)\frac{\gamma}{4[1-\Phi(z_1)]}dz_1$$

である。

3. 式（5）を満たす $Z_1,\ Z_2$ の確率は $z_1 \leq 0,\ z_2 \geq 0$ のとき

$z_1 \leq \Phi^{-1}(\alpha_1/2)$ の条件の下で、

$$\Phi(z_1)[1-\Phi(z_2)] \leq \frac{\gamma}{4} \tag{12}$$

が満たされる確率である。変形し

$$\Phi(z_2) \geq 1 - \frac{\gamma}{4\Phi(z_1)} \tag{13}$$

であるが、$z_2 \geq 0$ なので、$\Phi(z_2) \geq 1/2$ であり、

$$1 - \frac{\gamma}{4\Phi(z_1)} \leq \frac{1}{2}$$

の場合、つまり $\Phi(z_1) \leq \gamma/2$ の場合、即ち、$z_1 \leq \Phi^{-1}(\gamma/2)$ の場合、式（12）は常に満たされる。

それ以外、即ち、$z_1 \geq \Phi^{-1}(\gamma/2)$ の場合は、式（13）より、

$$z_2 \geq \Phi^{-1}(1 - \frac{\gamma}{4\Phi(z_1)})$$

が、式（12）を満たす条件である。ゆえに、求める確率は

$$P_{global-+} = \frac{1}{2}\int_{-\infty}^{\Phi^{-1}(\gamma/2)} \phi(z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\frac{\gamma}{4\Phi(z_1)}dz_1$$

である。

4. 式（5）を満たす $Z_1,\ Z_2$ の確率は $z_1 \leq 0,\ z_2 \leq 0$ のとき

$z_1 \leq \Phi^{-1}(\alpha_1/2)$ の条件の下で、

$$\Phi(z_1)\Phi(z_2) \leq \frac{\gamma}{4} \tag{14}$$

が満たされる確率である。変形し

$$\Phi(z_2) \leq \frac{\gamma}{4\Phi(z_1)} \tag{15}$$

であるが、$z_2 \leq 0$ なので、$\Phi(z_2) \leq 1/2$ であり、

$$\frac{\gamma}{4\Phi(z_1)} \geq \frac{1}{2}$$

の場合、つまり $\Phi(z_1) \leq \gamma/2$ の場合、即ち、$z_1 \leq \Phi^{-1}(\gamma/2)$ の場合、式（14）は常に満たされる。

それ以外、即ち、$z_1 \geq \Phi^{-1}(\gamma/2)$ の場合は、式（15）より、

$$z_2 \leq \Phi^{-1}(\frac{\gamma}{4\Phi(z_1)})$$

が、式（14）を満たす条件である。ゆえに、求める確率は

$$P_{global--} = \frac{1}{2}\int_{-\infty}^{\Phi^{-1}(\gamma/2)} \phi(z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\frac{\gamma}{4\Phi(z_1)}dz_1$$

である。

以上から、帰無仮説の下で最終的に、棄却域に落ちる確率、即ちタイプ I 過誤の確率は、

$$\begin{aligned}
P_{global} &= P_{global++} + P_{global+-} + P_{global-+} + P_{global--} = 4P_{global++} \\
&= \frac{1}{2}\int_{\Phi^{-1}(1-\gamma/2)}^{\infty} \phi(z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)} \phi(z_1)\frac{\gamma}{4[1-\Phi(z_1)]}dz_1 \\
&+ \frac{1}{2}\int_{\Phi^{-1}(1-\gamma/2)}^{\infty} \phi(z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)} \phi(z_1)\frac{\gamma}{4[1-\Phi(z_1)]}dz_1 \\
&+ \frac{1}{2}\int_{-\infty}^{\Phi^{-1}(\gamma/2)} \phi(z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\frac{\gamma}{4\Phi(z_1)}dz_1 \\
&+ \frac{1}{2}\int_{-\infty}^{\Phi^{-1}(\gamma/2)} \phi(z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\frac{\gamma}{4\Phi(z_1)}dz_1 \qquad (16)
\end{aligned}$$

となる。

ただし、上式の第 2,3,4,5 行目は、それぞれ $(z_1 \geq 0,\ z_2 \geq 0)$, $(z_1 \geq 0,\ z_2 \leq 0)$, $(z_1 \leq 0,\ z_2 \geq 0)$, $(z_1 \leq 0,\ z_2 \leq 0)$ に相当する項である。

しかし、ここで問題は $(z_1 \geq 0,\ z_2 \leq 0)$, $(z_1 \leq 0,\ z_2 \geq 0)$ の場合である。このような第一段階と第二段階の z の符号が異なる場合、研究者はその SNP を有意と判断するであろうか？おそらく、その SNP は有意性なしと判断されるであろう。そのように考える場合は、P 値積法における全体的なタイプ I 過誤の確率は、

$$\begin{aligned}
P_{global} &= P_{global++} + P_{global--} = 2P_{global++} \\
&= \int_{\Phi^{-1}(1-\gamma/2)}^{\infty} \phi(z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)} \phi(z_1)\frac{\gamma}{2[1-\Phi(z_1)]}dz_1 \\
&= \frac{\gamma}{2}(1 + \log\frac{\alpha_1}{\gamma})
\end{aligned}$$

となる。

次に、検出力を計算する。

対立仮説の下では、式（**??**）より次の統計量は

$$Z_1' = (\frac{X_m}{m_1} - \frac{Y_m}{m_2})/\sqrt{\frac{p_1(1-p_1)}{m_1} + \frac{p_2(1-p_2)}{m_2}} \sim N[(p_1-p_2)/\sqrt{\frac{p_1(1-p_1)}{m_1} + \frac{p_2(1-p_2)}{m_2}}, 1]$$

$$Z_2' = (\frac{X_m}{n_1} - \frac{Y_m}{n_2})/\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \sim N[(p_1-p_2)/\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, 1]$$

ただし、$X_m,\ Y_m$ は第一段階で、症例、および対照において、特定の（例えばリスク型）の遺伝子型を持つ個体数、$X_n,\ Y_n$ は第二段階で、症例、および対照において、特定の（例えばリスク型）の遺伝子型を持つ個体数である。$p_1,\ p_2$ はそれぞれ症例、対照の母集団における特定の（例えばリスク型の）遺伝子型の頻度であり、$m_1,\ m_2$ は第一段階での第一、第二集団のサンプルサイズ、$n_1,\ n_2$ は第二段階での第一、第二集団のサンプルサイズである。

上記の正規分布の平均を

$$\mu_1 = (p_1-p_2)/\sqrt{\frac{p_1(1-p_1)}{m_1} + \frac{p_2(1-p_2)}{m_2}}$$

$$\mu_2 = (p_1-p_2)/\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

とすると、検出力は

$$\begin{aligned}
W &= [1-\Psi(\mu_2,0)]\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\ z_1)dz_1 \\
&+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\ z_1)[1-\Psi(\mu_2,\ \Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\})]dz_1 \\
&+ \Psi(\mu_2,0)\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\ z_1)dz_1 + \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\ z_1)\Psi(\mu_2,\ \Phi^{-1}\{\frac{\gamma}{4[1-\Phi(z_1)]}\})dz_1 \\
&+ [1-\Psi(\mu_2,0)]\int_{-\infty}^{\Phi^{-1}(\gamma/2)}\psi(\mu_1,\ z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)}\psi(\mu_1,\ z_1)(1-\Psi\{\mu_2,\ \Phi^{-1}[1-\frac{\gamma}{4\Phi(z_1)}]\})dz_1 \\
&+ \Psi(\mu_2,0)\int_{-\infty}^{\Phi^{-1}(\gamma/2)}\psi(\mu_1,\ z_1)dz_1 + \int_{\Phi^{-1}(\gamma/2)}^{\Phi^{-1}(\alpha_1/2)}\psi(\mu_1,\ z_1)\Psi\{\mu_2,\ \Phi^{-1}[\frac{\gamma}{4\Phi(z_1)}]\}dz_1
\end{aligned} \tag{17}$$

ただし、$\psi(\mu,\ x),\ \Psi(\mu,\ x)$ は正規分布 $N(\mu,\ 1)$ のそれぞれ確率密度関数、累積分布関数を示し、$\phi(x),\ \Phi(x)$ はそれぞれ標準正規分布の確率密度関数、累積分布関数を示す。

また、上式の第 (1,2),3,4,5 行目は、それぞれ $(z_1 \geq 0,\ z_2 \geq 0),\ (z_1 \geq 0,\ z_2 \leq 0),\ (z_1 \leq 0,\ z_2 \geq 0),\ (z_1 \leq 0,\ z_2 \leq 0)$ に相当する項である。

しかし、対立仮説の下では（$p_1 > p_2$）実際には $(z_1 \geq 0,\ z_2 \leq 0)$、または、$(z_1 \leq 0,\ z_2 \geq 0)$ という結果が出る可能性は極めて低い。また、$p_1 > p_2$ の場合、$(z_1 < 0,\ z_2 < 0)$ の結果が出る可能性は更に極めて低い。従って、P 値積法の実際の検出力は、

$$\begin{aligned} W &= [1-\Psi(\mu_2,0)]\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\ z_1)dz_1 \\ &+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\ z_1)[1-\Psi(\mu_2,\ \Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\})]dz_1 \end{aligned} \tag{18}$$

である。

1.3 二つの段階から得られた統計量の関数として標準正規分布に従う統計量を算出する（joint 法)

第一段階で α_1 の有意水準で有意の SNP のみを第二段階でタイピングを行う。第一段階、第二段階のデータから得られる、母比率の差の推定量である統計量を標準正規化した統計量から標準正規分布に従う新たな統計量を定義する。

このような方法では実際の最終的なタイプ I の過誤の確率は、独立法の $\alpha_1\alpha_2$ より高い。帰無仮説が正しくても、第一段階では偶然有意の方向にかたよったデータが最終段階で再び用いられるので、第一段階と第二段階を独立に検定した場合よりタイプ I の過誤の確率が高くなるからである。しかし、真に関連した SNP であれば、第一段階でも有意に出やすいはずなので、検出力が向上する可能性も有る。

なお、joint 法は次の論文に発表されている (Skol AD et al. Nat Genet)。

第一回目のサンプルサイズを、症例と対照がそれぞれ $m_1,\ m_2$、第二回目のサンプルサイズも症例、対照が $n_1,\ n_2$ とする。

症例における頻度、p_1、対照における頻度、p_2 とすると、第**??**節、式（**??**）より、二つの群の割合の差の統計量は、第一段階、第二段階のデータについて、それぞれ

$$\begin{aligned} \frac{X_1}{m_1}-\frac{Y_1}{m_2} &\sim N(p_1-p_2,\ \frac{p_1(1-p_1)}{m_1}+\frac{p_2(1-p_2)}{m_2}) \\ \frac{X_2}{n_1}-\frac{Y_2}{n_2} &\sim N(p_1-p_2,\ \frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}) \end{aligned}$$

二つのランダム変数を標準正規化すると、

$$G_1 = \frac{X_1/m_1 - Y_1/m_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/m_1 + p_2(1-p_2)/m_2}} \sim N(0,1) \tag{19}$$

$$G_2 = \frac{X_2/n_1 - Y_2/n_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim N(0,1) \tag{20}$$

$p_1,\ p_2$ に相当する頻度が症例と対照で等しい p_c である帰無仮説の下では、

$$G_3 = \frac{X_1/m_1 - Y_1/m_2}{\sqrt{p_c(1-p_c)(1/m_1 + 1/m_2)}} \sim N(0,1)$$

$$G_4 = \frac{X_2/n_1 - Y_2/n_2}{\sqrt{p_c(1-p_c)(1/n_1 + 1/n_2)}} \sim N(0,1)$$

ここで p_c は対立仮説 $p_1 \neq p_2$ に対応する帰無仮説における、共通の $p_1 = p_2$ の値である。

しかし、これは実際の検定では観察できないので、観察データから推定することになる。第一段階、第二段階のデータから推定した p_c の値をそれぞれ $\hat{p_m},\ \hat{p_n}$ とすると、現実の検定で用いられる統計量は

$$G_3 = \frac{X_1/m_1 - Y_1/m_2}{\sqrt{\hat{p_m}(1-\hat{p_m})(1/m_1 + 1/m_2)}} \sim N(0,1)$$

$$G_4 = \frac{X_2/n_1 - Y_2/n_2}{\sqrt{\hat{p_n}(1-\hat{p_n})(1/n_1 + 1/n_2)}} \sim N(0,1) \tag{21}$$

ここで、それぞれの段階の症例と対照のサンプルサイズが同じ場合、即ち $m_1 = m_2 = m,\ n_1 = n_2 = n$ の場合のみを考えると

$$Z_1 = \frac{(X_1 - Y_1)/\sqrt{m}}{\sqrt{2\hat{p_m}(1-\hat{p_m})}} \sim N(0,1) \tag{22}$$

$$Z_2 = \frac{(X_2 - Y_2)/\sqrt{n}}{\sqrt{2\hat{p_n}(1-\hat{p_n})}} \sim N(0,1)$$

正規分布の再生成により、

$$\sqrt{m}Z_1 + \sqrt{n}Z_2 = \frac{X_1 - Y_1}{\sqrt{2\hat{p_m}(1-\hat{p_m})}} + \frac{X_2 - Y_2}{\sqrt{2\hat{p_n}(1-\hat{p_n})}} \sim N(0, n+m)$$

$$Z_v = \sqrt{\frac{m}{n+m}}Z_1 + \sqrt{\frac{n}{n+m}}Z_2$$

$$= \frac{X_1 - Y_1}{\sqrt{2(m+n)\hat{p_m}(1-\hat{p_m})}} + \frac{X_2 - Y_2}{\sqrt{2(m+n)\hat{p_n}(1-\hat{p_n})}} \sim N(0,1) \tag{23}$$

これが、joint 法で用いられている統計量である。

ここで、p_c の推定を第一段階と第二段階の全サンプルから推定された値とし、この値 $\hat{p_c}$ を、Z_1、Z_2 の両方に対して用いれば、

$$Z_u = \frac{(X_1 + X_2) - (Y_1 + Y_2)}{\sqrt{2(m+n)\hat{p_c}(1-\hat{p_c})}} \tag{24}$$

これは単に第一段階と第二段階のサンプルを併合して検定を行うときの統計量に過ぎない。なぜなら、**??**より、症例と対照のサンプルサイズが同じ n の場合の関連解析の統計量は

$$F_0 = \frac{X - Y}{n} \sim N[0, p_c(1-p_c)(\frac{2}{n})]$$

であり、これから

$$F_0/\sqrt{2p_c(1-p_c)/n} = \frac{X - Y}{\sqrt{2np_c(1-p_c)}} \sim N(0,1)$$

しかし、統計量（23）は標準正規分布に従うと仮定できるが、統計量（24）の分布は明らかでない。

今、帰無仮説の下で、第一段階の検定で有意となり、その上で有意となった SNP が第二段階でも有意となる確率、即ちタイプ I の過誤の確率を考える。Z_u も Z_1 も $N(0,1)$ に従うと考え、両側検定を考える。

第一段階の棄却域は、標準正規分布の累積分布関数を $\Phi(x)$ とすると、

$$z_1 \leq \Phi^{-1}(\alpha_1/2) \tag{25}$$

$$z_1 \geq -\Phi^{-1}(\alpha_1/2) \tag{26}$$

ただし、$\Phi^{-1}(x)$ は $\Phi(x)$ の逆関数である。第二段階の棄却域は

$$z_u \leq \Phi^{-1}(\alpha_u/2) \tag{27}$$

$$z_u \geq -\Phi^{-1}(\alpha_u/2) \tag{28}$$

式（25）または式（26）を満足し、しかも式（27）または式（28）を満足する場合に棄却域に落ちる（全体の棄却域）。

式（23）より、

$$z_u = \sqrt{\pi} z_1 + \sqrt{1-\pi} z_2$$

$$z_2 = \frac{z_u - \sqrt{\pi} z_1}{\sqrt{1-\pi}}$$

ただし、

$$\pi = \frac{m}{m+n}$$

とする。

式（27）と式（28）は、それぞれ次の条件と同じである。

$$z_2 \leq \frac{\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}} \tag{29}$$

$$z_2 \geq \frac{-\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}} \tag{30}$$

今、$z_u > 0$ に対応する棄却域のみを考えると（棄却域は $z_u > 0$ に対応するものと、$z_u < 0$ に対応する部分の確率を同じとする二つの部分に分けられる）が、$z_u > 0$ に対応する部分の確率 P_+ は、標準正規分布の確率密度関数を $\phi(x)$ とすると、

$$\begin{aligned} P_+ &= \int_{-\Phi^{-1}(\alpha_1/2)}^{\infty} \phi(z_1)\{1 - \Phi[\frac{-\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}]\} dz_1 \\ &\quad + \int_{-\infty}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\{1 - \Phi[\frac{-\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}]\} dz_1 \\ &\simeq \int_{-\Phi^{-1}(\alpha_1/2)}^{\infty} \phi(z_1)\{1 - \Phi[\frac{-\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}]\} dz_1 \end{aligned}$$

$z_u < 0$ に対応する部分の確率 P_- は、図 13-4 を参考に、

$$\begin{aligned} P_- &= \int_{-\infty}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\ \Phi[\frac{\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}] dz_1 \\ &\quad + \int_{-\Phi^{-1}(\alpha_1/2)}^{\infty} \phi(z_1)\ \Phi[\frac{\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}] dz_1 \\ &\simeq \int_{-\infty}^{\Phi^{-1}(\alpha_1/2)} \phi(z_1)\ \Phi[\frac{\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}] dz_1 \end{aligned}$$

最終的なタイプ I 過誤の確率は $P_+ + P_-$ であるが、$P_+ = P_-$ となるはずであり、

$$P = 2P_+ \simeq 2 \int_{-\Phi^{-1}(\alpha_1/2)}^{\infty} \phi(z_1)\{1 - \Phi[\frac{-\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}]\} dz_1 \tag{31}$$

が数理的に求めたタイプ I の過誤の確率である。

以上は帰無仮説の下で統計量が棄却域に落ちる確率（即ちタイプ I の過誤の確率）であるが、対立仮説の下では同じ統計量はどのような分布に従うであろうか。

式（19, 20) より対立仮説における $G_1,\ G_2$ の分布は

$$\frac{X_1/m_1 - Y_1/m_2}{\sqrt{p_1(1-p_1)/m_1 + p_2(1-p_2)/m_2}} \sim N(\frac{p_1-p_2}{\sqrt{p_1(1-p_1)/m_1 + p_2(1-p_2)/m_2}}, 1)$$
$$\frac{X_2/n_1 - Y_2/n_2}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim N(\frac{p_1-p_2}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}, 1)$$

$m_1 = m_2 = m,\ n_1 = n_2 = n$ の時

$$Z_1' \sim N(\frac{\sqrt{m}(p_1-p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}, 1)$$
$$Z_2' \sim N(\frac{\sqrt{n}(p_1-p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}, 1) \tag{32}$$

ここで、上式の $p_1,\ p_2$ の推定値を代入する（G_5 の $p_1,\ p_2$ は第一段階の症例と対照のサンプル、G_6 の $p_1,\ p_2$ は第二段階の症例と対照のサンプルで推定。

$$Z_1 \sim N(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2})}}, 1)$$
$$Z_2 \sim N(\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2})}}, 1) \tag{33}$$

従って、$Z_u = \sqrt{\pi}Z_1 + \sqrt{1-\pi}Z_2$ は（ただし、$\pi = m/(m+n)$)

$$Z_u \sim N(\frac{\sqrt{m+n}(p_1-p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}}, 1)$$

に従う。

対立仮説の下で第一段階で右、第二段階で右の棄却域に落ちる確率を P_{++}、第一段階で右、第二段階で左の棄却域に落ちる確率を P_{+-}、第一段階で左、第二段階で右の棄却域に落ちる確率を P_{-+}、第一段階で左、第二段階で左の棄却域に落ちる確率を P_{--} とすると、図 13-5 を参考に、

$$\begin{aligned} P_{++} &= \int_{-\Phi^{-1}(0,\alpha_1/2)}^{\infty} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2})}}, z_1) \\ &\times \{1 - \Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2})}}, \frac{-\Phi^{-1}(0,\alpha_u/2) - \sqrt{\pi}z_1}{\sqrt{1-\pi}}]\} dz_1 \\ P_{-+} &= \int_{-\infty}^{\Phi^{-1}(0,\alpha_1/2)} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2})}}, z_1) \\ &\times \{1 - \Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1}) + \hat{p_2}(1-\hat{p_2})}}, \frac{-\Phi^{-1}(0,\alpha_u/2) - \sqrt{\pi}z_1}{\sqrt{1-\pi}}]\} dz_1 \end{aligned}$$

$$
\begin{aligned}
P_{+-} &= \int_{-\Phi^{-1}(0,\alpha_1/2)}^{\infty} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, z_1) \\
&\times \Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, \frac{\Phi^{-1}(0,\alpha_u/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}]dz_1 \\
P_{--} &= \int_{-\infty}^{\Phi^{-1}(0,\alpha_1/2)} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, z_1) \\
&\times \{\Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, \frac{\Phi^{-1}(0,\alpha_u/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}]\}dz_1
\end{aligned}
\tag{34}
$$

ただし、$\phi(\mu, x)$, $\Phi(\mu, x)$ はそれぞれ、平均 μ、分散 1 の正規分布の確率密度関数と累積分布関数である。

ここで、P_{++} は $p_1 > p_2$ で、$p_1 - p_2$ が大きい場合、その他の確率よりはるかに大きく、多くの場合検出力は P_{++} のみで良い。しかし、p_1 と p_2 が接近している場合は P_{--} は無視できない。従って、検出力は

$$
\begin{aligned}
P_{++}+P_{--} &\simeq \int_{-\Phi^{-1}(0,\alpha_1/2)}^{\infty} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, z_1) \\
&\times \{1-\Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, \frac{-\Phi^{-1}(0,\alpha_u/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}]\}dz_1 \\
&+ \int_{-\infty}^{\Phi^{-1}(0,\alpha_1/2)} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, z_1) \\
&\times \{\Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, \frac{\Phi^{-1}(0,\alpha_u/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}]\}dz_1
\end{aligned}
$$

p_1 と p_2 が十分に離れている場合は

$$
\begin{aligned}
P_{++} &\simeq \int_{-\Phi^{-1}(0,\alpha_1/2)}^{\infty} \phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, z_1) \\
&\times \{1-\Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}}, \frac{-\Phi^{-1}(0,\alpha_u/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}]\}dz_1
\end{aligned}
$$

ここで、独立法、P 値積法、Joint 法の検出力の比較を行う。検出力を比較するためには、全体のタイプ I の過誤率をそろえなければならない。

まず、3 つの方法に共通の第一段階の有意水準を α_1 とする。次に、独立法における第二段階の有意水準を α_2 とする。P 値積法では、第一段階と第二段階の P 値の積に上限を設けるので、この値を γ とし、Joint 法では第一段階と第二段階での Z 統計量から得られる統計量 Z_u に対し有意水準を設

定するが、この有意水準を α_u とする。そうすると、全体のタイプ I の過誤の確率は、

独立法では

$$P_i = \frac{\alpha_1 \alpha_2}{2} \tag{35}$$

P 値積法では

$$P_m = \frac{\gamma}{2}(1 + \log \frac{\alpha_1}{\gamma}) \tag{36}$$

Joint 法ではこのように容易に書くことが出来ず、式（31）より、

$$P_j = 2P_+ \simeq 2 \int_{-\Phi^{-1}(\alpha_1/2)}^{\infty} \phi(z_1)\{1 - \Phi[\frac{-\Phi^{-1}(\alpha_u/2) - \sqrt{\pi} z_1}{\sqrt{1-\pi}}]\} dz_1$$

であった。ただし、ϕ, Φ はそれぞれ標準正規分布の確率密度関数、累積分布関数であり、$\pi = m/(m+n)$ である（m, n はそれぞれ第一段階、第二段階の症例のサンプルサイズであり、対照のサンプルサイズも等しいとする）。

まず、α_1, α_u, p_1, p_2, m, n に適当な値を入れ、P_j を計算する。次に、独立法において、これと同じ全体のタイプ I の過誤率を得るためには、$\alpha_2 = 2P_j/\alpha_1$ とすればよい。P 値積法においては、

$$P_j = \frac{\gamma}{2}(1 + \log \frac{\alpha_1}{\gamma})$$

を γ について解いて、

$$\gamma = -\frac{2P_j}{S[2P_j/(e\ \alpha_1)]} \tag{37}$$

ただし、S は、$y = S(x)$ において、$x = y\ e^y$ を満たす y、ただし、$-1 < y < 0$ を与える関数である。

そのようにして得られた、α_2, γ の値を用いて、それぞれ独立法、P 値積法の検出力を計算すればよい。

数理的な検出力の計算は、独立法では、式（3）のように、

$$\begin{aligned} W_i \quad = \quad & \simeq (1 - \Phi\{[\sqrt{\frac{2}{m} p_c(1-p_c)}\ \Phi^{-1}(1-\alpha_1/2) - p_1 + p_2] \\ & / \sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{m}}\}) \\ \times \quad & (1 - \Phi\{[\sqrt{\frac{2}{n} p'_c(1-p'_c)}\ \Phi^{-1}(1-\alpha_2/2) - p_1 + p_2] \\ & / \sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}\}) \end{aligned} \tag{38}$$

ただし、$p_c,\ p_c'$ はそれぞれ第一段階、第二段階での推定された帰無仮説の下でのリスク型遺伝子型の頻度である。これは、式（**??**）により計算される。

P 値積法の実際の検出力は、式（18）より、

$$\begin{aligned} W_m &= [1-\Psi(\mu_2,0)]\int_{\Phi^{-1}(1-\gamma/2)}^{\infty}\psi(\mu_1,\ z_1)dz_1 \\ &+ \int_{\Phi^{-1}(1-\alpha_1/2)}^{\Phi^{-1}(1-\gamma/2)}\psi(\mu_1,\ z_1)[1-\Psi(\mu_2,\ \Phi^{-1}\{1-\frac{\gamma}{4[1-\Phi(z_1)]}\})]dz_1 \end{aligned} \tag{39}$$

ただし、

$$\mu_1 = (p_1-p_2)/\sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{m}}$$
$$\mu_2 = (p_1-p_2)/\sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{n}}$$

、joint 法では

$$\begin{aligned} W_j &\simeq \int_{-\Phi^{-1}(0,\alpha_1/2)}^{\infty}\phi(\frac{\sqrt{m}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}},z_1) \\ &\times \{1-\Phi[\frac{\sqrt{n}(\hat{p_1}-\hat{p_2})}{\sqrt{\hat{p_1}(1-\hat{p_1})+\hat{p_2}(1-\hat{p_2})}},\frac{-\Phi^{-1}(0,\alpha_u/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}]\}dz_1 \end{aligned}$$

である。

図 13-6 に、全体のタイプ I の過誤率（global type I error rate）をそろえて、3 つの検定法を比較したグラフを示す。

検出力は Joint 法が最も高く、わずかに P 値積法が劣り、独立法はかなり劣る。しかし、この差は第一段階と第二段階のサンプルサイズが等しいあたりで最も大きく、第二段階のサンプルサイズが第一段階を大きく上回る条件では顕著ではなくなる。

しかし、このような数理的な検討は本当に正しいのであろうか。これを確かめるために、モンテカルロシミュレーションにより二段階絞込み法の全体のタイプ I の過誤率、及び検出力を検討しよう。

1.4 モンテカルロシミュレーション 2

まず、第一段階のサンプルサイズを $m_1,\ m_2$（症例、対照）、第二段階のサンプルサイズを $n_1,\ n_2$ とする。

SNP 情報による分類			
グループ　第一カテゴリー（リスク型の遺伝子型）		残り	合計
症例	m_{11}	$m_1 - m_{11}$	m_1
対照	m_{21}	$m_2 - m_{21}$	m_2
合計	$m_{11} + m_{21}$	$m_1 + m_2 - m_{11} - m_{21}$	$m_1 + m_2$

表 1: 第一段階で得られたデータの偶現表

1. 繰り返し回数を数えるために $i = 1$ とする。有意の検定数を数えるための数値 j を $j = 0$ に初期設定する。

2. 二つの二項分布 $B(m_1,\ p_1),\ B(m_2,\ p_2)$ から独立に $m_{11},\ m_{21}$ を取り出す。ただし、帰無仮説の時は $p_2 = p_1$。

3. 偶現表 1 を用いて、二群間の母比率の違いの検定、あるいは Pearson の χ^2 分布を用いた独立性の検定（どちらも同じ検定である）を行う。。

4. 独立法、および P 値積法の場合は、得られた P 値、P_1 が α_1 以下なら次に進む、それ以外は（2）に行く。Joint 法の場合は、偶現表 1 から次の統計量

$$z_1 \quad = \quad \frac{m_{11}/m_1 - m_{21}/m_2}{\sqrt{\hat{p}(1-\hat{p})(1/m_1 + 1/m_2)}}$$

を計算し（式 21 参照）、$|z_1| \geq \Phi^{-1}(1 - \alpha_1/2)$ なら（ただし、Φ^{-1} は標準正規分布の累積分布関数の逆関数）有意と考え、次に進む、それ以外は（2）に行く。

ただし、$\hat{p}$ は第一カテゴリーの帰無仮説の下での頻度の推定量であり、次の式により計算する。

$$\hat{p} = \frac{m_{11} + m_{21}}{m_1 + m_2}$$

二つの二項分布 $B(n_1,\ p_1),\ B(n_2,\ p_2)$ から独立に $n_{11},\ n_{21}$ を取り出す。ただし、帰無仮説の時は $p_2 = p_1$。

5. 表（1）の m を n に入れ換えた偶現表を用い、独立性の検定を行い、P 値を P_2 とする。

6. 独立法の場合は、$P_2 \leq \alpha_2$、P 値積法の場合は $P_1P_2 \leq \gamma$、Joint 法の場合は z_u を計算し（式 23 参照)、$|z_u| \geq \Phi^{-1}(1-\alpha_u/2)$ なら有意と考え有意性を判定し、有意なら有意の検定数を数えるために j を $j+1$ に更新する。

7. i が目的とする繰り返し回数 N になったら終了する、そうでなければ i を $i+1$ とし、(2）に戻る。

8. 繰り返し計算が終了したら、j/N を計算してタイプ I の過誤率の推定値（帰無仮説 $p_1 = p_2$ の場合)、あるいは検出力（対立仮説 $p_1 \neq p_2$ の場合）とする。

以上の方法で二段階絞込みによる関連解析の 3 つの検定法のタイプ I の過誤率と検出力を推定できる。図 13-7 に独立法と P 値積法について、数理的計算による検出力と、シミュレーションによる結果の比較を示す。いずれも極めて良く一致していることがわかる。

このようなモンテカルロシミュレーションの結果は、一定の条件下で、数理的解析の結果と非常に良く一致している。即ち、各サンプルサイズが十分大きく、p_1, p_2 がある程度大きく、各段階の検定が二群間の母比率の違いの検定、あるいは Pearson の χ^2 分布を用いた独立性の検定を用いた場合である。問題は、p_1, p_2 が小さい場合、あるいは各サンプルサイズが小さい場合である。このような場合は 2×2 の偶現表のセル中の期待度数が 5 以下になることがしばしばあり、二群間の母比率の違いの検定、あるいは Pearson の χ^2 分布を用いた独立性の検定を用いることが適当とは判断されない。そのような場合は Fisher の正確法が用いられることが多い。このように上記の一定の条件が満足されない場合は数理的なタイプ I の過誤率や検出力の推定は適当ではなく、シミュレーションを用いるべきである。

2　二段階絞込みによる SNP 探索の有意水準と検出力（制限がある場合）

第一段階と第二段階に分けて関連の有る SNP を探すとき、検出力を最大化するにはどのような条件を設定すればよいであろうか。

以下の考察では、特別に断らない限り症例・対照研究の偶現表の検定は Pearson の χ^2 分布に基づいた検定法による事にする。これは二群間の母比率の違いを正規分布を用いて行う検定法と同じものである（第**??**節）。

しかし、実際に研究を行う場合、色々な制限があり、自由に研究計画を立てられるわけではない。まず、全タイピングコストが固定されているとき、どのように研究を進めるべきかを考えよう。症例と対照は人数に制限が無いことにする。第一段階、第二段階それぞれで症例と対照のサンプルサイズが同じ場合のみを考える。

2.1 全タイピングコストが制限されている場合（独立法）

今、第一段階でタイピングする全 SNP 数 s_1 と全タイピングコスト h_t が与えられているとする。また、第一段階と第二段階のそれぞれの SNP あたり、一人当たりのタイピングコストを、$c_1,\ c_2$ とする。

第一段階と第二段階にかけるコストをどのように振り分ければ最も効率的であろうか。第一段階のサンプルサイズが決まれば（第一段階での総 SNP 数は固定されているので）、第一段階に必要なコストは決まり、従って、第二段階で使用できるコストは決まる。ここで第二段階でのサンプルサイズが決まれば、第二段階でのタイピング SNP 数が決まる。従って、この問題で必要な二つの変数は、第一段階での症例のサンプルサイズ N_1 と第二段階での症例のサンプルサイズ N_2 である。対照のサンプルサイズは症例と同じなので、第一段階では $2N_1$ 人、第二段階では $2N_2$ 人がタイピングされることになる。

第一段階では s_1 個の SNP についてタイピングされ、1 つの SNP の一人あたりのタイピングコストは c_1 なので、第一段階での総タイピング費用は

$$H_1 = 2N_1 c_1 s_1 \tag{40}$$

第二段階では $2N_2$ 人がタイピングされ、1 つの SNP のタイピングコストは c_2 なので、第二段階でタイピング可能な SNP 数は

$$S_2 = \frac{H_2}{2N_2 c_2} = \frac{h_t - 2N_1 c_1 s_1}{2N_2 c_2}$$

ただし、

$$N_1 \leq \frac{h_t}{2c_1 s_1}$$

第一段階で有意となり、第二段階に送られる SNP のほとんどすべては関連の無い SNP である。帰無仮説の下で、第一段階で有意となる確率が第一段階での有意水準 α_1 である。従って、

$$\alpha_1 = \frac{S_2}{s_1} = \frac{h_t - 2N_1 c_1 s_1}{2N_2 c_2 s_1} \tag{41}$$

上式で変数は N_1、および N_2 のみであり、残りは定数である。従って、$N_1,\ N_2$ が決まれば α_1 は決まる。α_1 が決まれば全体のタイプ I の過誤率から、それぞれの方法に応じて $\alpha_2,\ \gamma,\ \alpha_u$ が決まる。例えば、全体のタイプ I の過誤率を P_g とすると、独立法の場合は、$\alpha_2 = 2P_g/\alpha_1$ であり、P 値積法の場合は、$P_g = \gamma[1 + \log(\alpha_1/\gamma)]/2$ を解いて得られる γ の値である（式 37 参照）。Joint 法の場合は少し面倒であるが、式（37）の数値積分より、与えられた P_j を満足する α_u の値を計算する。

このように、α_1 に加え、独立法の場合は α_2、P 値積法の場合は γ、joint 法の場合は α_u が定まれば、数理的に、あるいはモンテカルロシミュレーションにより検出力を計算できる。

動ける $N_1,\ N_2$ の範囲から、検出力が最大となる $N_1,\ N_2$ を求めればよい。

例えば、$s_1 = 50,000,\ c_1 = 1,\ c_2 = 5,\ h_t = 50,000,000$ の場合を考える。また、対立仮説の $p_1 = 0.3,\ p_2 = 0.2$（オッズ比 1.71）とする。求める全体のタイプ I 過誤の確率は、Bonferroni の修正法を採用し、$0.05/s_1 = 10^{-5}$ とする。検定は独立法を用いることにする。

式（41）より、

$$\alpha_1 = \frac{50000000 - 2N_1 \times 50000}{2N_2 \times 5 \times 50000} = \frac{500 - N_1}{5N_2} \tag{42}$$

これを、（N_1 を m に、N_2 を n に変えた後）、式（38）に代入し、検出力を $N_1,\ N_2$ の二つの変数の関数として表す。この場合、独立法なので、第二段階での有意水準は $\alpha_2 = 2P_i/\alpha_1$ により定まる（式 35）。ただし、P_i は全体のタイプ I 過誤の確率。

図 13-8 に N_1 のいくつかの値に対し、検出力 W_i のグラフを示す。

二変数を動かすことにより最大値を求めれば $N_1 = 418,\ N_2 = 1078$ の時、検出力は最大で 0.808 となり（図 13-8）、この時、$\alpha_1 = 0.0152,\ \alpha_2 = 0.00131$ となる、最初の SNP 数、50,000 個のうち、選択され第二段階に渡される割合は α_1 と等しく、その数は $S_2 = 761$ となる。

2.2 全タイピングコストが制限されている場合（P 値積法）

全タイピングコストが制限されているが、サンプルサイズには制限が無く、独立法ではなく P 値積法により検定を行う場合を考える。この場合も、第一段階の検定を α_1 で行うところまでは同じである。しかし、第二段階での検定の有意水準を、独立法では $\alpha_2 = 2P_i/\alpha_1$ で決めたが、P 値積法では、$P_m = \gamma[1 + \log(\alpha_1/\gamma)]/2$ を満足する γ を求め（式 36）、第一段階と第二段階の P 値の積が γ 以下であれば有意とする。そして、与えられた $\alpha_1,\ \gamma$ の下での検出力は式（39）で表された。

例として、上記の独立法の場合と同じ条件を考える。$s_1 = 50,000,\ c_1 = 1,\ c_2 = 5,\ h_t = 50,000,000$ である。また、対立仮説の $p_1 = 0.3,\ p_2 = 0.2$（オッズ比 1.71）とする。求める全体のタイプ I 過誤の確率は、Bonferroni の修正法を採用し、$0.05/s_1 = 10^{-5}$ とする。検定は独立法ではなく、P 値積法を用いることにする。

式（42）により、α_1 が $N_1,\ N_2$ の関数として表され、$P_m = 10^{-5} = \gamma[1 + \log(\alpha_1/\gamma)]/2$ より γ を求める。さらに、これを、（N_1 を m に、N_2 を n に変えた後）、式（18）に代入し、検出力を計算する。図 13-9 にいくつかの N_1 の下での N_2 と検出力の関係を示す。二変数 $N_1,\ N_2$ を動かすことにより最大値を求めれば $N_1 = 420,\ N_2 = 1007$ の時、検出力は最大で 0.816 となり（図 13-9）、この時、$\alpha_1 = 0.0159,\ \gamma = 2.00 \times 10^{-6}$ となる、最初の SNP 数、50,000 個のうち、選択され第二段階に渡される割合は α_1 と等しく、その数は $S_2 = 794$ となる。即ち、P 値積法の方が少し検出力が高い。

2.3 全タイピングコストと、全サンプルサイズの両方に制限がある場合（独立法）

しかし、この条件では症例、対照とも全体のサンプルサイズ 1496（独立法の場合）、あるいは 1427（P 値積法の場合）が必要であるが、これほどの人数が集まるかどうか不明である。症例、対照それぞれの全体のサンプルサイズが N_t と制限されている場合は、異なった条件設定となる。この場合は、第一段階のサンプル数 N_1 のみが変数となり、第二段階のサンプル数 N_2 は $N_2 = N_t - N_1$ により決まる（ただし、$N_t > 1496$ の場合は上記のように $N_1 = 418,\ N_2 = 1078$ が最適である。全体のサンプルサイズに制限がある場合、第一段階の有意水準は、

$$\alpha_1 = \frac{h_t - 2N_1 c_1 s_1}{2(N_t - N_1)c_2 s_1} \tag{43}$$

である。

前の例と同じ、$s_1 = 50,000,\ c_1 = 1,\ c_2 = 5,\ h_t = 50,000,000$ の場合を考える。また、対立仮説の $p_1 = 0.3,\ p_2 = 0.2$（オッズ比 1.71）とする。求める全体のタイプ I 過誤の確率を 10^{-5} とし、検定は独立法を用いることにする。ただし、症例、対照ともサンプルサイズの制限 1,000 がある。また、サンプルサイズは不足しており、第一段階と第二段階で準備された $N_t = 1000$ のサンプルはすべて使い切るとする。独立法の検出力を示す、式（38）は（m を N_1 に、n を $N_2 = N_t - N_1$ と入れ換える）N_1 のみの関数となり、容易に最大化することが出来る。図 13-10 に N_1 と検出力の関係を示す。検出力を最大化する条件は、$N_1 = 317,\ N_2 = 683,\ \alpha_1 = 0.0536,\ \alpha_2 = 0.0000135$ であり、第一段階の結果を見て、第二段階に、割合 0.0536 の SNP、即ち、1698 個の SNP が回される。得られた検出力は 0.638 である。

2.4 全タイピングコストと、全サンプルサイズの両方に制限がある場合（P 値積法）

さらに、P 値積法を用いた場合の最適化について考察しよう。

P 値積法の場合も、前の例と同じ、$s_1 = 50,000,\ c_1 = 1,\ c_2 = 5,\ h_t = 50,000,000$ の場合を考える。また、対立仮説も $p_1 = 0.3,\ p_2 = 0.2$（オッズ比 1.71）と、前回と同様とする。求める全体のタイプ I 過誤の確率を 10^{-5} とし、検定は P 値積法を用いることにする。また、症例、対照ともサンプルサイズの制限は 1,000 である。また、サンプルサイズは不足しており、第一段階と第二段階で準備された $N_t = 1000$ のサンプルはすべて使い切るとする。

N_1 が決まれば $N_2 = N_t - N_1$ と、式（43）より、α_1 が決まり、P 値積法の場合の全体のタイプ I の過誤率の式（36）より、式（37）を参考にし、γ が決まる（$P_j = 10^{-5}$ である）。α_1 と γ が決まれば、P 値積法の場合の検出力の式（39）より、検出力が決まる。このようにして、N_1 を変化させて、検出力が最大になる条件を探せば良い。図 13-11 に N_1 と検出力の関係を示す。それは、$N_1 = 439,\ N_2 = 561,\ \alpha_1 = 0.0217$ であり、第二段階に回される SNP 数は 1088 個である。この時の検出力は 0.729 である。これは独立法による最大検出力 0.638 よりかなり高い。また、第二段階に回される SNP 数が検定法により異なるのも特徴的である。

以上のように、条件により最適の $\alpha_1,\ \alpha_2$ はかなり異なり、従って、第二段階にまわされる SNP の

割合も異なる。当然、最終的な検出力も異なる。

条件をさまざま変更し、まず $N_1,\ N_2$ に制限を設けず検出力の推定を行い、最適のサンプル数が得られない場合はサンプルサイズを制限して検出力の推定を行うべきである。

3　図の説明

図 13-4　症例・対照研究の二段階法を joint 法で行った場合のタイプ 1 過誤率の計算

本文を参照の事

図 13-5　症例・対照研究の二段階法を joint 法で行った場合の検出力の計算

本文を参照の事

図 13-6　症例・対照研究の二段階法を 3 つの異なった手法を用いた場合の検出力の比較

条件は、第一段階の有意水準 $\alpha_1 = 0.05$, 症例と対照における頻度 $p_1 = 0.3,\ p_2 = 0.2$、第一段階の標本サイズは症例、対照とも 400 とした。3 つの手法による最終的なタイプ 1 過誤の確率が 10^{-6} となるように調製した。横軸は、第二段階における標本サイズであり、症例と対照で同じである。

図 13-7　二段階法の検出力計算のためのシミュレーションと数理的方法の比較

二段階法の症例・対照研究の検出力をシミュレーション（Sim）と数理的方法（Math）を用いて計算し比較した。検定法は独立法（Ind）と P 値積法（Multi）。条件は症例と対照における頻度 $p_1 = 0.3,\ p_2 = 0.2$、第一段階の有意水準は $\alpha = 0.05$、第一段階での標本サイズは症例、対照ともに 200、第二段階での標本サイズを横軸に変化させた。全体の第 1 種の過誤の確率を 0.00005 にした。

図 13-8　二段階の症例・対照研究のための検出力を最大化するための条件の最適化（独立法）

検定法は独立法により、タイピングコストに制限があるが、標本サイズは無制限に与えられた場合の検出力の検討。症例と対照の頻度を $p_1 = 0.3,\ p_2 = 0.2$、第一段階のタイピングス SNP 数を 50,000、第一段階、第二段階のタイピングコストを 1SNP あたり、1,5、全タイピングコスト 5×10^7、最終的なタイプ 1 の過誤率を 0.00001 とした。第一段階と第二段階の標本サイズは独立に変化させうるが、全タイピングコストの制限により、第一段階から第二段階に送る候補 SNP の数が変化する。

図 13-9　　二段階の症例・対照研究のための検出力を最大化するための条件の最適化（P 値積法）

検定法は P 値積法を用い、タイピングコストには制限があるが人数は無制限に用いることができる場合の最適条件を検討した。タイピングコストに制限があるが、標本サイズは無制限に与えられた場合の検出力の検討。症例と対照の頻度を $p_1 = 0.3,\ p_2 = 0.2$、第一段階のタイピングス SNP 数を 50,000、第一段階、第二段階のタイピングコストを 1SNP あたり、1,5、全タイピングコスト 5×10^7、最終的なタイプ 1 の過誤率を 0.00001 とした。第一段階の標本サイズ（症例と対照で同サイズ）$N1$ と第二段階の標本サイズの両方を変化させた。

図 13-10　二段階の症例・対照研究のための検出力を最大化するための条件の最適化（タイピングコストと全人数の両方に制限がある場合、独立法）

検定法は独立法を用い、タイピングコストと人数の両方に制限がある場合の最適条件を検討した。症例と対照の頻度を $p_1 = 0.3,\ p_2 = 0.2$、第一段階のタイピングス SNP 数を 50,000、第一段階、第二段階のタイピングコストを 1SNP あたり、1,5、全タイピングコスト 5×10^7、最終的なタイプ 1 の過誤率を 0.00001 とした。第一段階の標本サイズ（症例と対照で同サイズ）$N1$ と第二段階の標本サイズの和を 1,000 に制限した。

第 13-11　二段階の症例・対照研究のための検出力を最大化するための条件の最適化（タイピングコストと全人数の両方に制限がある場合、P 値積法）

検定法は P 値積法を用い、タイピングコストと人数の両方に制限がある場合の最適条件を検討した。症例と対照の頻度を $p_1 = 0.3,\ p_2 = 0.2$、第一段階のタイピングス SNP 数を 50,000、第一段階、第二段階のタイピングコストを 1SNP あたり、1,5、全タイピングコスト 5×10^7、最終的なタイプ 1 の過誤率を 0.00001 とした。第一段階の標本サイズ（症例と対照で同サイズ）$N1$ と第二段階の標本サイズの和を 1,000 に制限した。

参考文献

[1] Pratt SC, Daly MJ, Kruglyak L (2000) Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. Am J Hum Genet 66: 1153-1157

[2] Lange K, Westlake J, Spence MA (1976) Extensions to pedigree analysis: III. Variance components by the scoring method. Ann Hum Genet 39, 485-491

[3] Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc 82:605-610

[4] Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535-543

[5] Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19: 1-17

家系図による個人のリスク（発症確率）の推定

付録

1　家系図による個人のリスク（発症確率）の推定

1.1　比較的単純な例でのリスクの推定

まず、極めて単純な家系によるリスクの推定を行う。この疾患は常染色体性完全優性の遺伝病とする。また、疾患関連座位は既に明らかであり、リスクをもたらすアレル A、非リスクに関連するアレル a が存在するとする。検査によりこの座位の遺伝子型が観察できるとし、完全優性なので $AA,\ Aa$ の遺伝子型であれば必ず発症し、aa であれば発症することは無いとする。この場合は、発症する確率は AA または Aa の遺伝子型となる確率と一致する。

このような遺伝形式に従う疾患を持つ家系があるとする（図 1a)。分離の法則により、子 (?) の遺伝子型の確率は $Aa,\ aa$ がそれぞれ $1/2$ である。従って、子の発症確率は $1/2$ である。次に、図 1b の場合、遺伝形式は常染色体の完全劣性であるとする。疾患関連座位の遺伝子型 $Aa,\ Aa$ の両親の子 (?) の遺伝子型の確率は $AA,\ Aa,\ aa$ が $1/4, 1/2, 1/4$ である。AA のときのみ発症するので、子の発症確率は $1/4$ である。

1.2　疾患関連座位はわかるが特定の個人の遺伝子型が不明の場合

しかし、一般にこのような単純な場合は稀である。遺伝形式が完全優性、完全劣性の場合ばかりではない（不完全浸透、または不完全優性)。更に、年齢により発症率が異なる場合もある。家系内の特定の個人の遺伝子型が不明の場合もある。更には、疾患関連座位の遺伝子型が不明で、それと連鎖する座位の遺伝子型のみがわかる場合もある。

まず、疾患座位は明らかであり、遺伝形式も常染色体性の完全優勢、または完全劣性であるが、特定の個人の遺伝子型が不明の場合を考える。特定の個人の遺伝子型が明らかでないことは、死亡、不明、診断の拒否などで実際に存在する。

特に、親の遺伝子型が不明の場合は、子の遺伝子型から親の遺伝子型を推定するという方法を用いる必要がある。即ち、結果（子）をもとに原因（親）を推定するという手法である。この時、確率の概念を用いるため理解が困難になる。親（原因）をもとに子（結果）の確率を考える過程はもともとの確率の定義に合っているが、結果をもとに原因の確率を考えることは困難をともなう。しかし、ベイズの定理を用いれば、そのような考察が可能である。

しかし、ベイズの定理は現実世界に適用するには注意が必要な定理である。図 2 の家系について実際にベイズの定理を適用してみる。遺伝的データは因果が明確なので、ベイズの定理の問題点と利点が理解しやすい。

常染色体性完全優性の遺伝病について、家系図で個体 D の遺伝子型がわからないので、その推定を行いたいとする。完全優性なので遺伝子型がわかれば発症するかどうかは完全にわかる。しかし、個体 D の消息が不明であり遺伝子型も発症の有無もわからない。

個体 D については二種類の情報が存在する。第一の情報は、D の父母（この家系では祖父母という事になる）の遺伝子型である。それは明確にわかっており、どちらも Aa である（従って発症している）。第二は、D の子供三人の遺伝子型がわかっており、それはすべて aa である。この二種類の情報は D にとっては完全に異なるものである。第一の情報は D の遺伝子型の原因に関する情報であり、第二の情報は D の遺伝子型の結果に関する情報である。

祖父母の遺伝子型については、確かにこれが異なれば D の遺伝子型に大きな影響を与えるであろう。従って、祖父母の遺伝子型情報を D の遺伝子型推定に用いるのは合理的である。しかし、D の子供の遺伝子型についてはどうであろうか。D に子供が居ようが居まいが、子供の遺伝子型が何であろうと D の遺伝子型に変わりはないはずである。そのような考え方からは、D の子供の遺伝子型は無視して、祖父母の遺伝子型のみを考え、D の遺伝子型の確率分布は $AA(\frac{1}{4}), Aa(\frac{1}{2}), aa(\frac{1}{4})$ で解析終了とする立場も理解できる。即ち、メンデルの分離の法則を応用しただけである。

しかし、よくよく考えると、D の子供の遺伝子型の情報を無視する事はあまりにも不合理である事がわかる。D に aa の子供ができているという事は D の遺伝子型が AA である可能性はすでに 0 である事を意味する。もし、D に Aa の遺伝子型の子供が居るとすると（実際には居ないが）D の遺伝子型が aa である可能性は消失する。また、D の子供が 100 人いて（こんなことはありえないが）、そのすべての遺伝子型が aa だとすると、D の遺伝子型が aa である可能性がだんぜん高くなるであろう。

そもそも D の遺伝子型は既に決まっているのである。確かに、D の遺伝子型の結果から、原因を推定する事は納得がいかないようにも思うが、ここは我慢して考える事にする。

結果から原因の確率を推定するためには標本空間を極めて厳密に定義する必要がある。図 2 の例では、一つの実験とは「遺伝子型が与えられている三人の創始者から、すべての非創始者にいきわた

るようにメンデルの分離の法則に従って、A または a のアレル伝達を行う」と定義し（この定義は因果の関係を乱していないことに注意）、その結果の集合を標本空間とする。これらの結果のうち、孫の一人でもその遺伝子型が aa で無いものは、観察データに合致しない事になる。観察データに合致した結果だけを集めて、それを出来事とし、その中で D の遺伝子型が Aa である出来事の割合が事後確率である。即ち、

$$\begin{aligned} & P(S_{Aa} \mid obs) \\ = \quad & P(S_{Aa})P(obs \mid S_{Aa})/(P(S_{AA})P(obs \mid S_{AA}) + P(S_{Aa})P(obs \mid S_{Aa}) + P(S_{aa})P(obs \mid S_{aa})) \end{aligned}$$

obs は観察データ（に合致した出来事）、S_{gen} は D の遺伝子型が gen である出来事を示す（図 3 参照）。

事前確率は、$P(S_{Aa}) = \frac{1}{2}, P(S_{AA}) = \frac{1}{4}, P(S_{aa}) = \frac{1}{4}$ であり、条件付確率は、$P(obs \mid S_{Aa}) = (\frac{1}{2})^3, P(obs \mid S_{AA}) = 0, P(obs \mid S_{aa}) = 1^3$ なので、事後確率は $P(S_{Aa} \mid obs) = \frac{1}{5}$

同様に、$P(S_{AA} \mid obs) = 0, P(S_{aa} \mid obs) = \frac{4}{5}$ である。完全優性なので発症確率は AA または Aa となる確率に一致し、この親 D が発症する確率は 1/5 である。親の遺伝子型のみを利用した解析（事前確率）が、子の遺伝子型を利用することにより変化した（事後確率）ことがわかる。

これは、このような実験を無限回繰り返した時、その中で観察データに合致する結果の中で、D の遺伝子型が AA, Aa, aa である割合に一致する。

ここで、問題となるのが事前確率である $P(S_{AA}), P(S_{Aa}), P(S_{aa})$ である。この問題では、これらはすべてメンデルの分離の法則に基づいている事がわかる。これが、ベイズの定理を適用した場合の信頼性を支えている。もし、このような事前確率の信頼性が保証されないような例にベイズの定理を適用すると、信頼できない結果しか得られないであろう。ここに、現実世界の問題にベイズの定理を適用する事の問題点があるのである。例えば、D の父母の遺伝子型が不明であるとするとどうであろう。S_{AA}, S_{Aa}, S_{aa} しか可能性が無く、それに関して情報が全く無い場合に、それらを等確率と考える事は妥当であろうか。もちろん、それは妥当ではない。

ベイズの定理は、結果の情報から原因を確率で推定するものである。現実世界にそれを適用した場合の異様さは遺伝の問題を考えると良く理解できると思う。

この疾患が完全優性の遺伝形式を取るときは、発症するかどうかは遺伝子型を見ればわかる。しかし、不完全優性（不完全浸透）の場合はそうではない。ただ、それが疾患関連座位である事が明らか

な場合は浸透率の概念を用いれば発症確率を計算できる。

遺伝子型 $AA,\ Aa,\ aa$ に対応する浸透率を、それぞれ $q_{AA},\ q_{Aa},\ q_{aa}$ とすると、個人 D の発症確率は

$$P(S_{AA}|obs)q_{AA} + P(S_{Aa}|obs)q_{Aa} + P(S_{aa}|obs)q_{aa} = \frac{1}{5}q_{Aa} + \frac{4}{5}q_{aa} \tag{1}$$

例えば $q_{Aa} = 0.7,\ q_{aa} = 0$ とわかっていれば、発症確率は 0.14 である。

1.3 個人の遺伝子型がわからず、浸透率がわかる場合

図 4a は極めて稀な常染色体性不完全優性の遺伝病であり、個人 1 のみが発症している。この疾患の浸透率が 0.7 であることがわかっているとする。疾患関連座位は不明で、遺伝子型も不明である。

個人 1 は発症しており、極めて稀な遺伝病なので遺伝子型は Aa であり、近親結婚ではないとすると、配偶者の遺伝子型は aa である。従って、個人 2 の遺伝子型は Aa と aa の可能性がある。もし個人 2 の遺伝子型が aa であれば個人 3 が発症する可能性はゼロに近いが、個人 2 が Aa であると個人 3 は発症の可能性がある。

個人 2 について事前の遺伝子型の確率は $Aa,\ aa$ がともに 1/2 である。しかし、発症していないので（Aa の場合、その確率は 0.3、aa の場合 1）、Aa の事後確率は 0.23 である（計算は以下の通り）。

$$\frac{0.3 \times \frac{1}{2}}{0.3 \times \frac{1}{2} + \frac{1}{2}} \simeq 0.23$$

以上より、個人 3 については、Aa の確率は $0.23 \times 0.5 \simeq 0.115$ であり、発症の確率は $0.115 \times 0.7 \simeq 0.081$ である。これは、個人 3 がまだ産まれていない場合、産まれた後で発症する可能性と解釈することができ、妊娠前に結果の確率を計算する方法として用いる事ができる。

1.4 浸透率が年齢などで変化する場合

ここで問題は、浸透率が年齢により変化するという問題である。

次のような年代で発症する確率がわかっているとする。

20 代未満 0%, 20 代 25%, 30 代 35%, 40 代 20%, 50 代 10%, 60 代 10%

即ち 60 代までには必ず発症する。ということは、20 代終わりまでに発症しない確率は 75%、30 代終わりまでに発症しない確率は 40%、40 代終わりまでに発症しない確率は 20%、50 代終わりまでに発症しない確率は 10%である。

今、図 4b のような家系があり、個人 2, 3 は発症しておらず、年齢がそれぞれ 59 歳, 15 歳とする。個人 2 の遺伝子型の事前確率は $Aa,\ aa$ が 1/2 であるが、59 歳で発症していないので、事後確率を計算すると

$$\begin{aligned} P(Aa|obs) &= \frac{P(Aa)P(obs|Aa)}{P(Aa)P(obs|Aa)+P(aa)P(obs|aa)} = \frac{1/2\times 0.1}{1/2\times 0.1+1/2\times 1} \simeq 0.091 \\ P(aa|obs) &= 1-P(Aa|obs) = 0.909 \end{aligned}$$

ただし、ここで個人 3 は 15 歳なので遺伝子型が Aa であったとしても発症の確率は 0 であり、個人 2 の遺伝子型推定の役に立たないことに注意が必要である。以上より、個人 3 の遺伝子型が Aa である確率は $0.091\times 0.5 \simeq 0.045$ である。

しかし、浸透率は 10 歳ごとに急に変化するのではなく、連続的に変化すると考えられる。例えば、AA の個体については図 5a のように年齢と伴に発症率が漸増するとする。グラフ $F_{AA}(x)$ は特定の年齢 x で発症している患者の割合を示すとする。特定の年齢で発症する割合は $F_{AA}(x)$ を x で微分した関数 $f_{AA}(x)$ で表される。$f_{AA}(x)$ は確率密度関数、$F_{AA}(x)$ は累積分布関数に相当する関数である。そうすると、年齢 x においても発症しない確率は 1-$F_{AA}(x)$ となり、年齢 x における微小期間 Δx の間に発症する確率は $f_{AA}(x)\Delta x$ である。ただし、Δx は微小期間とする。このような考察によりリスクの計算が可能になる。この疾患では aa の遺伝子型では発症しないが、Aa では AA より低い確率で発症し、その年齢ごとの発症確率が $F_{Aa}(x)$ で表されるとする（図 5b)。確率密度関数は $f_{Aa}(x)$ である。

図 6 の家系は稀な常染色体性の遺伝病の家系であるとし、遺伝子型 $AA,\ Aa$ はそれぞれ図 5a,b のような年齢ごとの発症曲線に従い、遺伝子型 aa では発症しないとする。表 1.4 に個人 2, 3 の遺伝子型の候補とその事前確率、個人 2 の表現型（疾患あり）の確率、個人 3 の表現型（疾患無し）の確率を示す。以上より、ベイズの定理を用いて、3 の遺伝子型が aa である事後確率は

$$P(aa|obs) = \frac{1/4\times f_{Aa}(51)\Delta x}{1/4\times f_{Aa}(51)\Delta x+1/2\times f_{AA}(51)\Delta x[1-F_{Aa}(25)]+1/4\times f_{Aa}(51)\Delta x[1-F_{Aa}(25)]}$$

遺伝子型（2-3）	遺伝子型の事前確率	2 の表現型の確率	3 の表現型の確率
$AA - Aa$	$1/2$	$f_{AA}(51)\Delta x$	$1 - F_{Aa}(25)$
$Aa - Aa$	$1/4$	$f_{Aa}(51)\Delta x$	$1 - F_{Aa}(25)$
$Aa - aa$	$1/4$	$f_{Aa}(51)\Delta x$	1

$$= \frac{1/4 \times f_{Aa}(51)}{1/4 \times f_{Aa}(51) + 1/2 \times f_{AA}(51)[1 - F_{Aa}(25)] + 1/4 \times f_{Aa}(51)[1 - F_{Aa}(25)]}$$

以上の場合は、創始者の遺伝子型が与えられたので比較的考察や容易であった。例えば祖父母の遺伝子型がわからない場合は、リスクアレル A の集団における頻度や、個人の表現型をもとに創始者の遺伝子型の事前確率を計算しなければならない。

1.5 疾患関連座位から既知の組み換え割合の位置にある座位の遺伝子型による推定

これまでの考察では、疾患関連座位の遺伝子型のみを問題にした。しかし、場合によっては疾患関連座位の遺伝子型が観察できず、それから既知の θ だけの組み換え割合の位置のマーカー座位の遺伝子型のみが観察できる場合もある。その場合は、組み換えの可能性を考慮したリスクの計算を行う必要がある。

図 6b は稀な常染色体性不完全優性の疾患の家系とする。しかし、疾患座位はわかっていないものの、それと組み換え割合 θ の距離で連鎖するマーカーがわかっており、そのマーカー座位の遺伝子型が観察できる。図 6b の遺伝子型はそのマーカー座位の遺伝子型を示す。疾患関連座位については稀な遺伝病なので、図の中の罹患者の遺伝子型は Aa である（A がリスクアレルとする）。また $Aa,\ aa$ の浸透率は 0.9 と 0 とする。

個人 1 から 2 に渡されるハプロタイプは $A - 2$ であるが、個人 2 から 3 へは $a - 2,\ A - 2$ の両方の可能性があり、それぞれが確率 $\theta,\ 1 - \theta$ である。そして、前者なら 3 は病気にならず、後者なら 3 は 0.9 の確率で病気になる。従って、個人 3 の発症確率は $0.9(1 - \theta)$ である。もし、個人 3 がある年齢まで発症していないとすると、その情報を考慮して発症確率を更に低く計算しなければならないことはもちろんである。

1.6　家系情報からの発症リスクの計算の一般的方法

以上のように、家系情報から個人の発症リスクを計算するためには色々な情報を考慮しなければならない。一般的に、そのような計算法を行えば、間違いなく個人の発症確率を計算できるであろうか。

個人的リスクの計算はパラメトリック連鎖解析で行った計算法と極めて類似している。疾患関連座位と連鎖し、情報が得られるすべてのマーカー座位の遺伝子型、更には個人の表現型や浸透率などを考慮し個々の結果の確率は次のようになる（「遺伝統計学入門」の中の式*.*），

$$\begin{aligned} P(h,w,s,\Phi) &= (\prod_{i}^{f}\prod_{j}^{L}\prod_{k}^{2} p_{ijk}) \times 2^{-2n} \times [\prod_{i=1}^{2n}\prod_{j=1}^{L-1} \theta_j^{r_{ij}}(1-\theta_j)^{1-r_{ij}}] \\ &\quad \times \prod_{j=1}^{3} q_j^{v_{j0}}(1-q_j)^{v_{j1}}. \end{aligned} \tag{2}$$

ただし、変数の説明は「遺伝統計学入門」の本文を参照の事。

このような結果の一つ一つに、注目の個人の疾患関連座位の遺伝子型（$AA,\ Aa,\ aa$）が対応している。また、このような結果一つ一つが観察データに合致するか確認できる。以上の情報から、観察データのもとでの注目の個人の疾患関連座位の遺伝子型（$AA,\ Aa,\ aa$）の事後確率を計算すればよい（図 3 参照）。疾患座位の遺伝子型の事後確率が計算できれば、その個人の発症確率は浸透率のデータから計算可能である。

図の追加

図8-10 Log quantile-quantile (QQ) *P*-value plot
網羅的関連解析で、3つのモードによる解析の最小の*P*を採用した場合

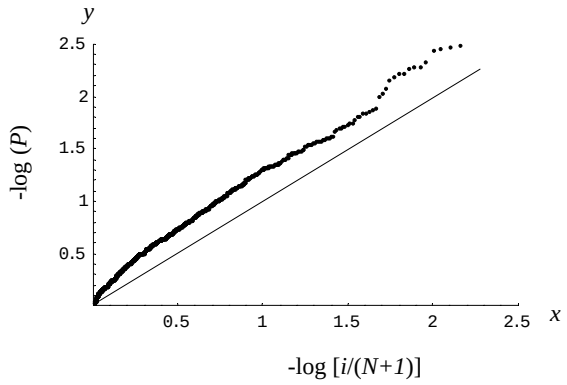
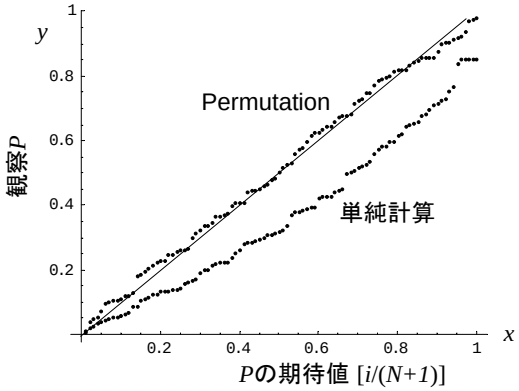
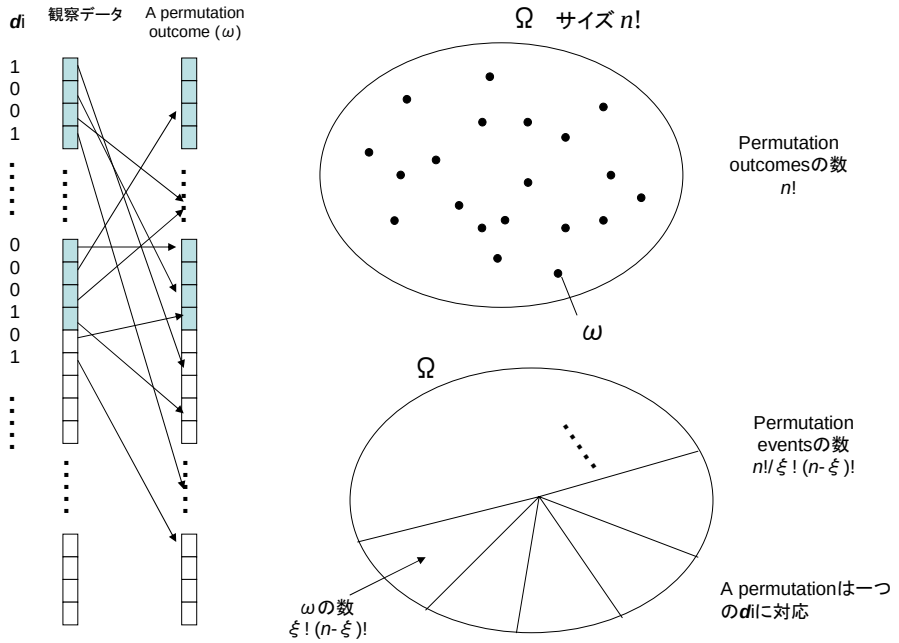


図8-10 網羅的関連解析で、3つのモードによる解析の最小の*P*を採用した場合
Permutationによる*P*との比較





	Genotype 1	Genotype 2	
Disease	a	b	ξ
Nondisease	c	d	$n-\xi$
	n_1	n_2	

$\max(0,\ \xi + n_1 - n) \leqq a \leqq \min(\xi,\ n_1)$

j 座位について、$\min(\xi,\ n_1) - \max(0,\ \xi + n_1 - n) + 1$個の偶現表が対応

偶現表 (a,b,c,d)に対し、${}_{n1}C_a \times {}_{n2}C_{\xi - a} = n_1!\, n_2!/[a!\,(n_1 - a)!(\xi - a)!(n_2 - \xi + a)!]$個のpermutation eventsが対応

偶現表(a,b,c,d)の確率は$n_1!\, n_2!\ \xi!\,(n-\xi)!/[n!\ a!\,(n_1 - a)!(\xi - a)!(n_2 - \xi + a)!]$
これは周辺度数を固定したとき、aを唯一の変数とする超幾何分布

一つのpermutation eventの確率は$\xi!\,(n-\xi)!/n!$

一つのpermutation eventに$\xi!\,(n-\xi)!$個のpermutation outcomesが対応

一つのpermutation outcomeの確率は$1/n!$

図 補遺-6 ケースとコントロールにおける割合の反転

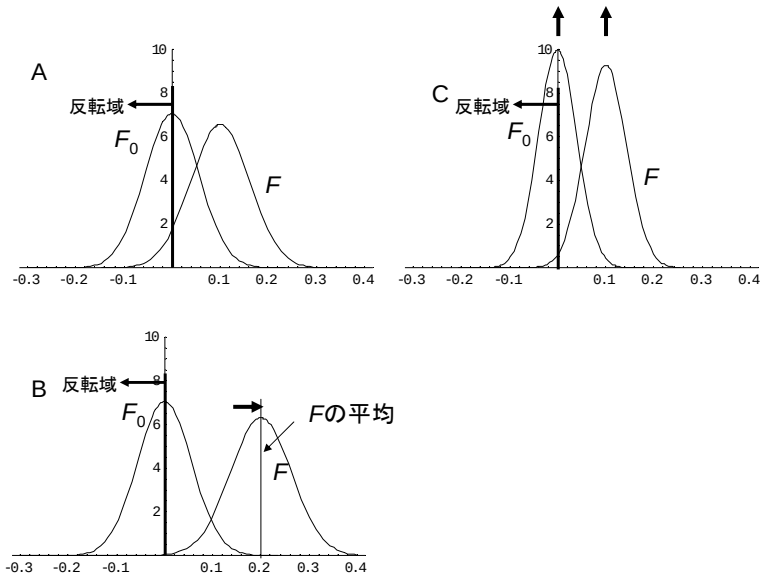


図 補遺-7 ケース・コントロール研究で割合が反転する確率

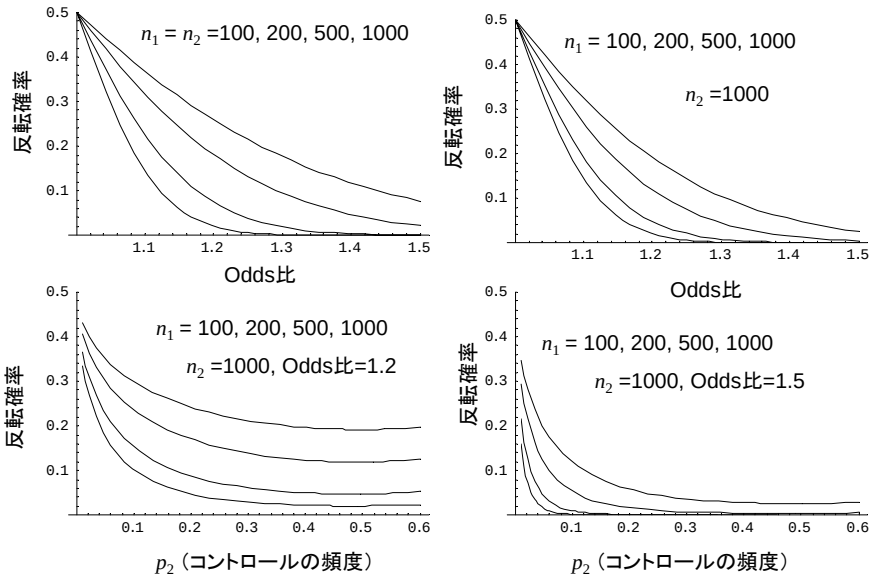


図1

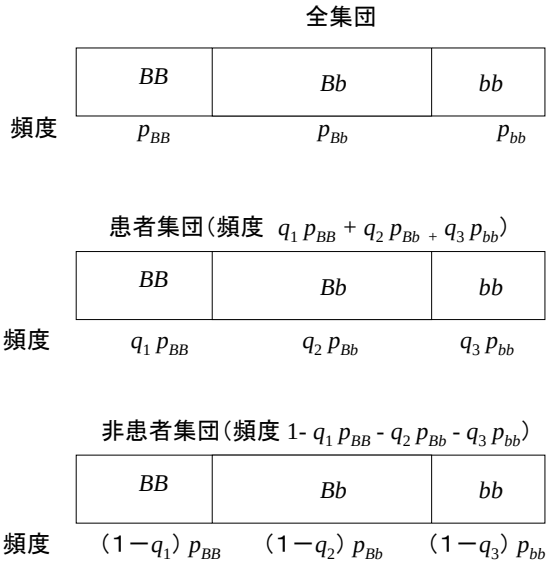


図 2　量的表現型値の分布と浸透率の関係

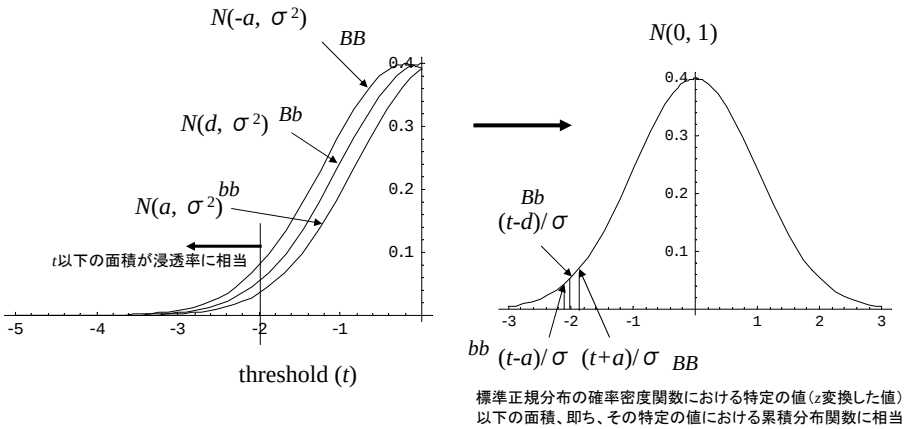
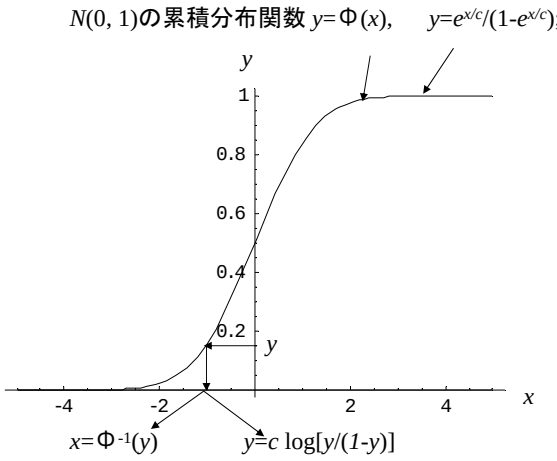
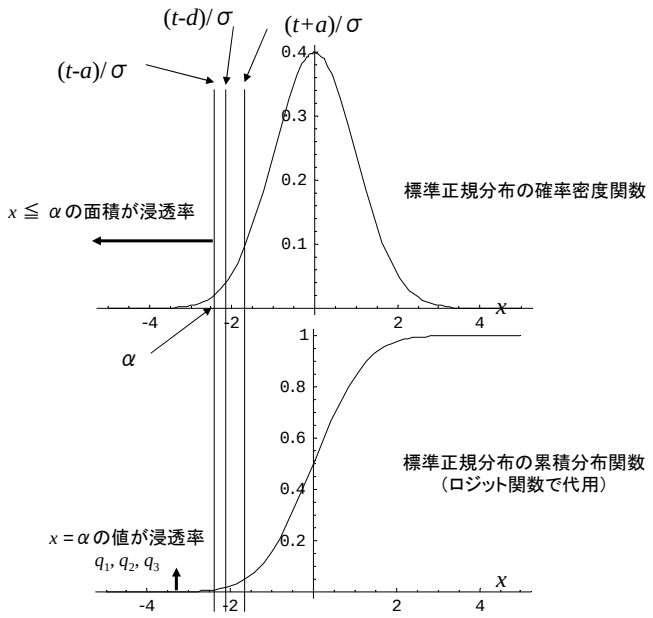


図3

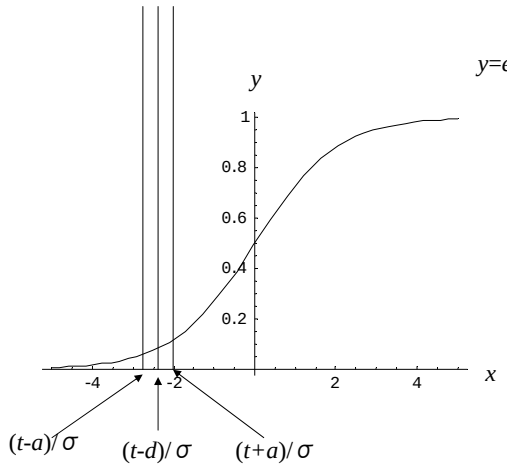


ロジット関数（c=0.61475の時）と平均0、標準偏差1の正規分布の累積分布関数は極めて良く似ている（重なっているので違いがわからない）

図3



アレルが相加的に作用する場合、*Bb*に対応する*x*軸の値が*BB*, *bb*の平均



$y=e^{x/a}/(1-e^{x/a})$

逆関数

$x=1/a \log[y/(1-y)]$

アレルが相加的に作用する場合、**ロジット関数を利用し、浸透率(*y*)を使って**
**オッズ *y*/(*1*-*y*) を計算すると、*Bb*に対応するオッズの対数は*BB, bb*の平均になる

図 13-4

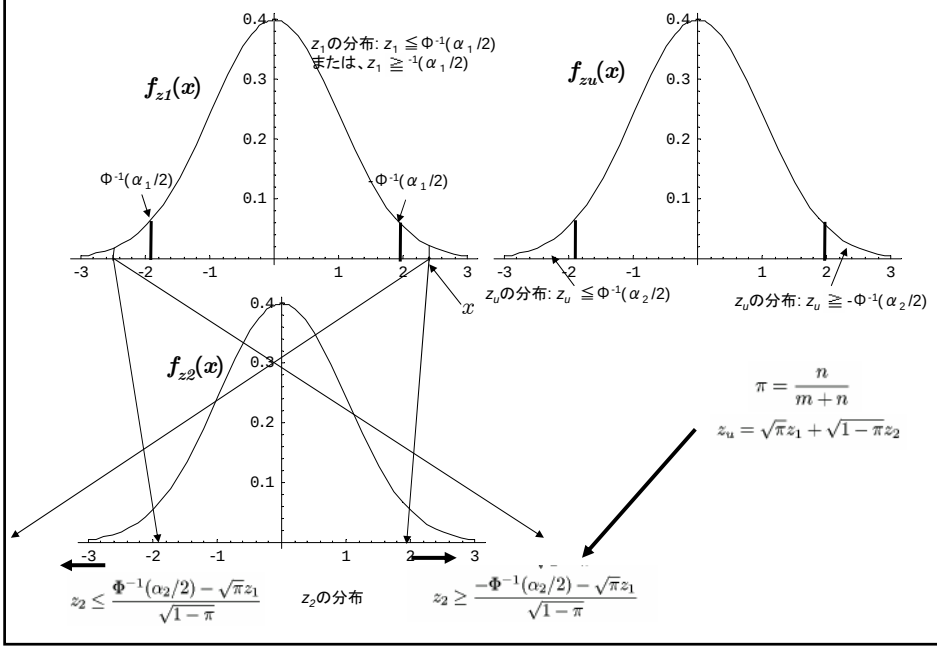


図 13-5

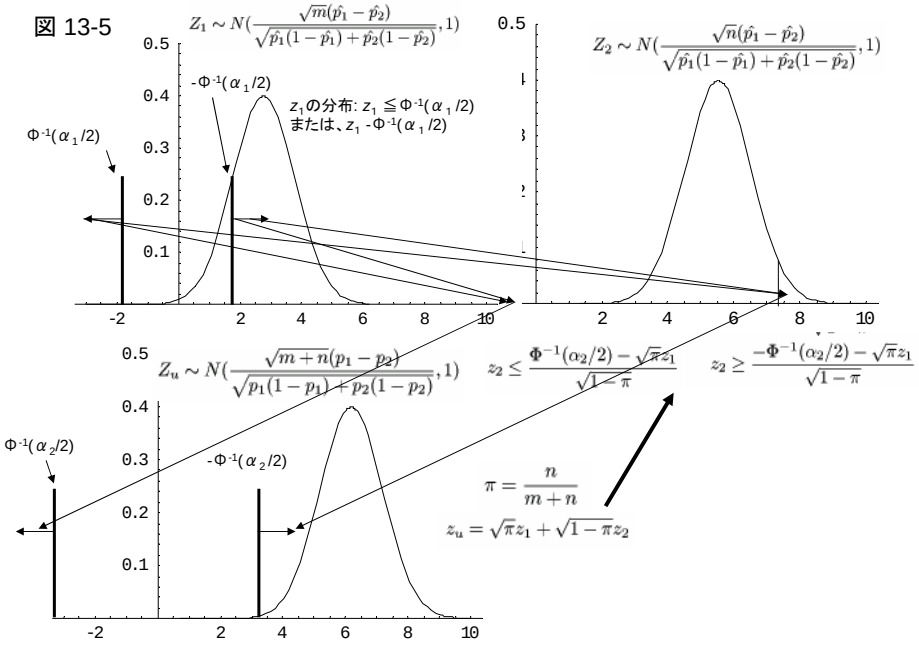

$Z_1 \sim N(\frac{\sqrt{m}(\hat{p}_1-\hat{p}_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)}},1)$
$Z_2 \sim N(\frac{\sqrt{n}(\hat{p}_1-\hat{p}_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)}},1)$
$Z_u \sim N(\frac{\sqrt{m+n}(p_1-p_2)}{\sqrt{p_1(1-p_1)+p_2(1-p_2)}},1)$
$-\Phi^{-1}(\alpha_1/2)$
$\Phi^{-1}(\alpha_1/2)$
z_1の分布: $z_1 \leqq \Phi^{-1}(\alpha_1/2)$
または、$z_1 - \Phi^{-1}(\alpha_1/2)$
$\Phi^{-1}(\alpha_2/2)$
$-\Phi^{-1}(\alpha_2/2)$
$z_2 \leq \frac{\Phi^{-1}(\alpha_2/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}$
$z_2 \geq \frac{-\Phi^{-1}(\alpha_2/2)-\sqrt{\pi}z_1}{\sqrt{1-\pi}}$
$\pi = \frac{n}{m+n}$
$z_u = \sqrt{\pi}z_1 + \sqrt{1-\pi}z_2$

図 13-6

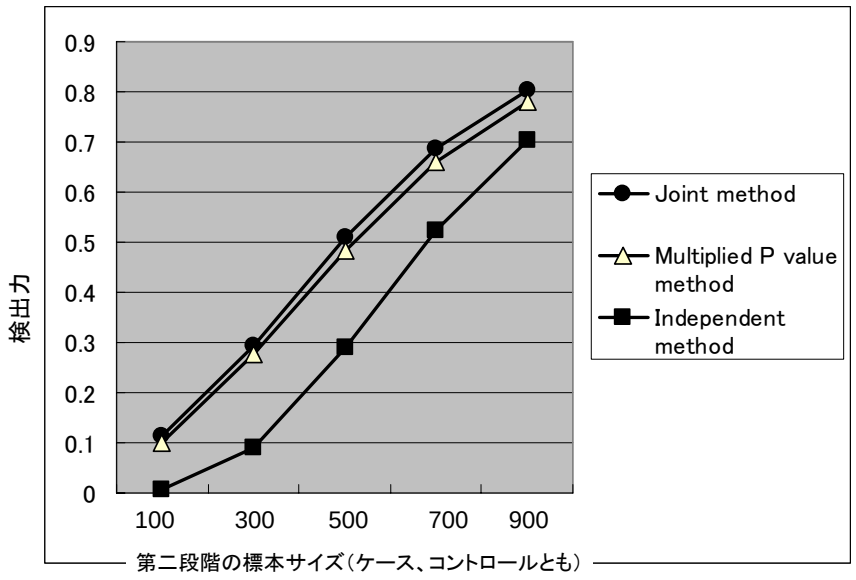

検出力
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
300
500
700
900
第二段階の標本サイズ（ケース、コントロールとも）
Joint method
Multiplied P value method
Independent method

図 13-7

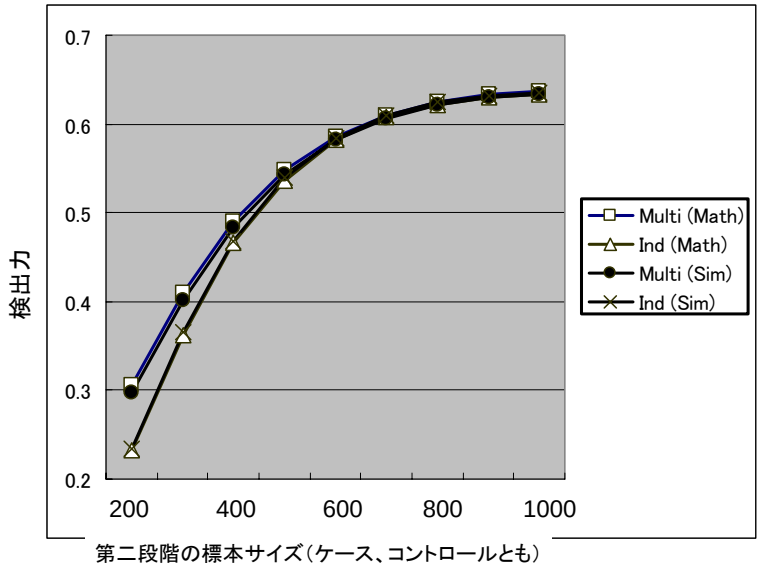

0.7
0.6
0.5
0.4
0.3
0.2
検出力
200
400
600
800
1000
第二段階の標本サイズ（ケース、コントロールとも）
Multi (Math)
Ind (Math)
Multi (Sim)
Ind (Sim)

図 13-8

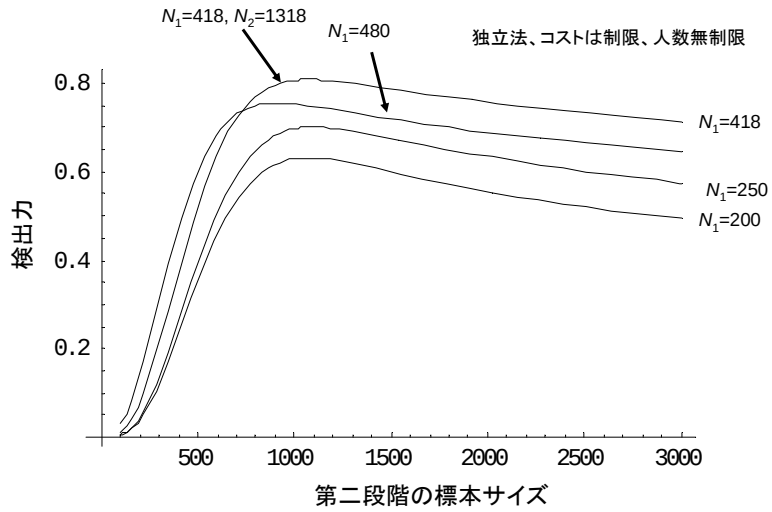
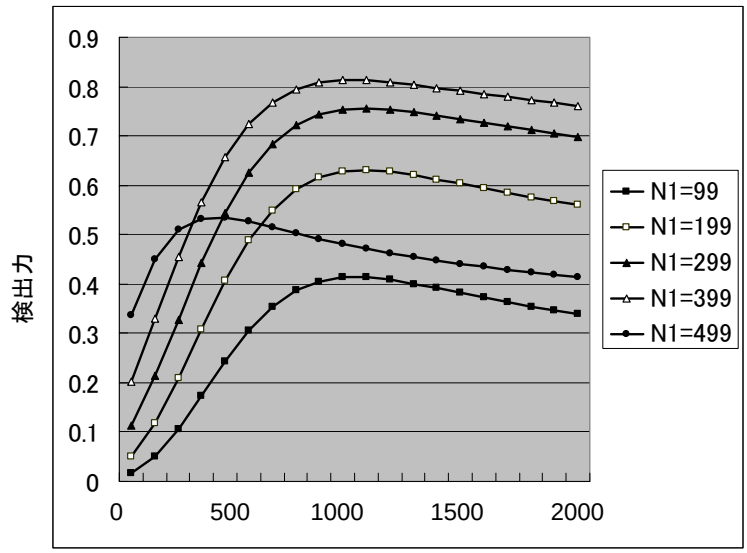

N_1=418, N_2=1318
N_1=480
独立法、コストは制限、人数無制限
0.8
0.6
0.4
0.2
検出力
N_1=418
N_1=250
N_1=200
500
1000
1500
2000
2500
3000
第二段階の標本サイズ

図 13-9

P値積法、コストは制限、人数無制限

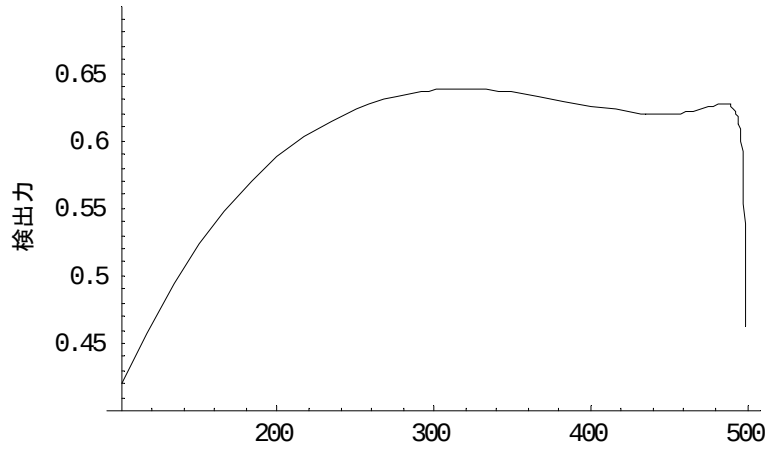


OR=1.71429

第二段階の標本サイズ

図 13-10

独立法、N1+N2=1000の制限



OR=1.71429

第二段階の標本サイズ

図 13-10

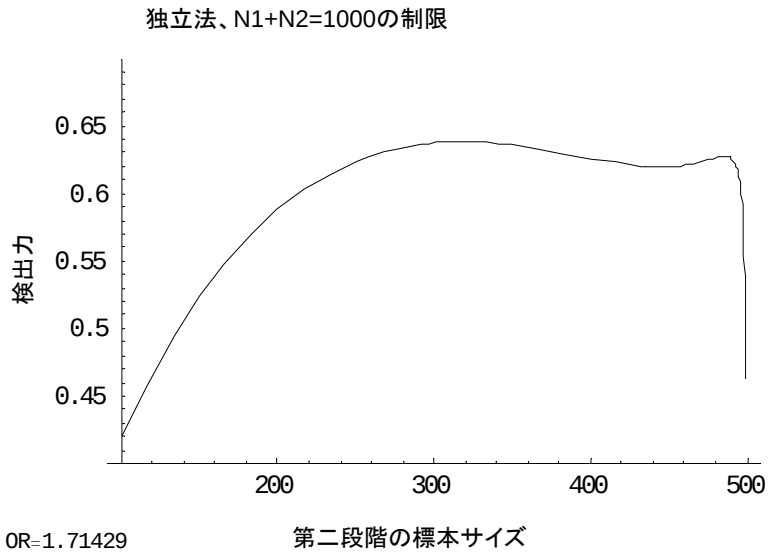
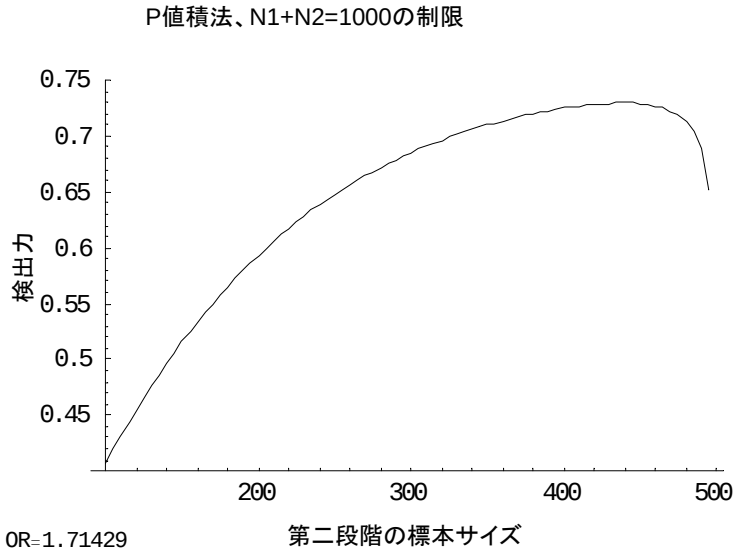

独立法、N1+N2=1000の制限
検出力
0.65
0.6
0.55
0.5
0.45
200
300
400
500
OR=1.71429
第二段階の標本サイズ

図 13-11


P値積法、N1+N2=1000の制限
検出力
0.75
0.7
0.65
0.6
0.55
0.5
0.45
200
300
400
500
OR=1.71429
第二段階の標本サイズ

遺伝統計学入門・補遺集

2008 年 1 月 7 日

著者　鎌谷直之
発行所　株式会社スタージェン
東京都台東区蔵前 4-31-10 蔵前オラシオンビル 9 階
電話　03-5835-2137