法数学勉強会

山田　亮

2013 年 8 月 15 日

はじめに

この文書は京都大学医学部法医学講座にて 2010 年から始まった法数学勉強会で提供した話題をまとめたものです。同勉強会は DNA 多型を用いた個人識別を取り巻く話題の数学的側面を勉強しようと言う会で、私は DNA 多型に関する統計学を遺伝疫学分野にて専門にしていることから、法医学講座の玉木敬二教授にお誘いいただいて参加してきたものです。多型についてはわかっていても法医学分野での問題設定・情報の解釈と活用は、私にとってまったく初めてのテーマであり、参加のたびに多くのことを勉強させていただいております。このように門外漢として出発した私の文書ですので、内容は的外れであったり、基本的な誤解があったりするかもしれません。その点はあらかじめご容赦いただきたいと思います。それとは逆に門外漢であるが故に、法医学分野の常識にとらわれない考えもあるかもしれず、それについては、ポジティブな側面があるかもしれないという温かい眼で見ていただき、多少なりとも参考になることがあれば幸いです。本文書は章立てをしてはいますが、全体の統一をとるための作業をほとんどしておりません。複数の章で同様の記述が繰り返されることがあるかもしれませんが、ご容赦ください。

目次

第Ⅰ部

法数学勉強会の記録

第 1 章

仮説空間、確率と尤度 2010 年 8 月

勉強会で使用したスライドです。以下の文章と勉強会での話の内容には違う点もありますが、参考までにリンクを張っておきます。

1.1 はじめに

DNA 鑑定で一番多い状況は、「この DNA 試料の型が仮説 X と仮説 Y とのどちらによりよく合致するか」を尤度比で数値化するものだと思われます。この章では、特定２仮説の比較ではなく、比較するべき仮説が２つとは限らない場合とはどういう場合なのかを問題にします。

1.2 確率と尤度

1.2.1 条件付き確率

ある仮説の下で、ある事象が起きる確率を $Pr(\mathbf{x}|\theta)$ と書きます。$\mathbf{x}$ はいろいろな事象を、θ は仮説の条件を表します。$\mathbf{x}$ と θ との間の縦棒"|"は $A|B$ で『B が成り立つという条件での A』と読みます。従って $Pr(\mathbf{x}|\theta)$ は『θ が成り立っているときに $\mathbf{x}$ が起きる確率』と読みます。条件が付いた確率なので、条件付き確率と呼びます。

1.2.2 起きうる事象と生起確率

$\mathbf{x}$ は太字で書いてあります。これは $\mathbf{x}$ は一つだけでなくたくさん (場合によっては $0 \leq x \leq 1$ のようにある範囲について数限りない値のことを想定する場合もあります。

DNA 鑑定では、個人が持つジェノタイプ (複数マーカーの組合せジェノタイプ) の一つ一つが $\mathbf{x}$ の一つ一つの要素に対応します。だれしも、何かしらのジェノタイプを持ちますし、あるジェノタイプを持ったら、他のジェノタイプは持てませんから、$\mathbf{x}$ のすべてについて確率を計算してその値を足しあわせれば、1 になります。

$$\sum_{x \in \mathbf{x}} Pr(\mathbf{x}|\theta) = 1$$

と書いたり ($\mathbf{x}$ が離散的な場合)

$$\int_{x \in \mathbf{x}} Pr(\mathbf{x}|\theta) dx = 1$$

と書いたり ($\mathbf{x}$ が連続的な場合) します。

1.2.3　起きた事象と尤度

DNA 鑑定で、ある試料についてジェノタイプが実験的に確定していれば、その特定のジェノタイプ x_0 が問題になります。今、ある条件 (その試料がある個人由来である、という仮説) を取れば、この確率は $Pr(x_0|\theta)$ となります。これは確率の式そのものですが、ジェノタイプとして観測し終えたものを想定しているので、呼び方を変えることにして、『θ の下での尤度』と呼びます。確率と尤度の関係については 10 章も参考にしてください。

1.3　２つの仮説の比較

今、試料の由来主は２人のみに限られるとすると、その２人のそれぞれが x を観測する尤度としては $Pr(x|\boldsymbol{\theta_1}), Pr(x|\boldsymbol{\theta_2})$ の２つが考えられます。尤度が高い方が、「真の由来主」だろうと思えます。その「真の由来主」らしさを表す方法の一つが尤度比です。

$$\frac{Pr(x|\boldsymbol{\theta_1})}{Pr(x|\boldsymbol{\theta_2})}$$

は、２人目に比べて１人目が真の由来主らしさは何倍かを数字にしたものです。
別の方法で「真の由来主」らしさを数字にする方法があります。

$$\frac{Pr(x|\boldsymbol{\theta_1})}{Pr(x|\boldsymbol{\theta_1}) + Pr(x|\boldsymbol{\theta_2})}$$

とする方法です。尤度の相対的比率です。この場合は最小で 0、最大で 1 です。尤度比は「●倍」で大小を表します。差がなければ「１倍」です。尤度の相対的比率は単位はあり

ません。「0.9」なら 0.9 です。差が無い時は「0.5」です。どちらも数字が違うだけで、本来の意味は変わりません。

1.4　仮説は２つとは限らない〜DNA 鑑定においても〜

DNA 試料の由来主が３人以上いる場合は、仮説も３つ以上になります。その場合、尤度比はあくまでも、２つの仮説の尤度の比なので、n 個の仮説に対して $\frac{n(n-1)}{2}$ 個の尤度比が計算できます。もし、尤度比を使って、ある特定の仮説の真偽を考えようとすれば、この $\frac{n(n-1)}{2}$ 個の数字についての判断が必要になります。簡単な考え方としては、$\frac{n(n-1)}{2}$ 個の尤度比の最小値でさえもある値より大きい、というような基準でしょう。たくさんの尤度比を判断にあたってどのように使うかについてのコンセンサスは存在していないと思います。
一方、尤度の相対的比率であれば、

$$\frac{Pr(x|\theta_{\mathbf{1}})}{\sum_i Pr(x|\theta_{\mathbf{i}})}$$

と、２仮説の比較の場合と同じ枠組みです。

1.5　立場が違うときには、棄却したい仮説が対立する

DNA 鑑定では２つの異なる立場からの主張が対立することがあると思います。片方の主張 (主張 X) は、『試料のジェノタイプは A さん以外の人に由来すると仮定すると、あまりにも尤度が低いので、A さん以外の人に由来するという仮定は間違っていると思います。A さん由来のはずです』というものです。もう片方の主張 (主張 Y) は、『試料が A さん由来である可能性もありますが、誰かひとりでも A さん以外の誰かがいて、その人由来であるという尤度が十分に低くなければ、A さんであると断定することはできないはずです』

というものです。主張 X では、A さんである、という仮説以外のすべての仮説について確率を出して、その総和確率 (場合によっては事前確率で重みづけした総和) が十分に小さいことを示せばよいです。他方、主張 Y では、たくさんの仮説があるなかで、ただ一つでも、A さんとの違いが十分大きくない人がいることを示せばよいです。

主張 X では総和を問題にしていますが、多くの場合は、現実的な計算上の理由から、平均値を使います。主張 Y では、特定の１つだけについて問題にします。

もし、特定の１つの仮説だけは、「A さんである」という仮説と十分な差がなく、でも

それは 1 つの仮説だけで、そのほかの大量の仮説は「A さんである」という仮説と十分な差がある、という場合には、主張 X と主張 Y とはどちらも成り立ってしまいます。

　たくさんの仮説 (DNA 鑑定ではたくさんの候補者) があって、そのうちわけが不均一であるときに、このようなどっちつかずが起きます。そのような不均一の最たるものは、一卵性双生児同胞の存在です。

したがって、仮説をどのように設定するかは個々の主張にも影響しますし、複数の主張でどっちつかずになるような場合にも大きく影響してきます。

第 2 章

尤度比検定 2011 年 2 月

2.1 はじめに

前回は確率と尤度と仮説に関する内容でした。また、尤度比は２つの仮説のもっともらしさを比較する指標であることも扱いました。今回は、その尤度比を使った検定の話です。同じく尤度比を使いますが、使い方が少し違うので、その違いを押さえることを目的とします。

2.2 モデルと尤度

前回書いた通り、ある仮説の下である事象が観察される確率を、観察事象が得られたあとで問題にして、仮説のもっともらしさを表す数値とみなしたとき、それを尤度と言います。仮説のことをモデルと言うこともあります。少し言葉のニュアンスが違うとすると、「モデル」という場合には、変数を導入してその変数を使って確率・尤度が計算できる、という印象が強いかもしれません。たとえば、「サイコロを３回振ったら、4,2,4 が出た」という事象に対して、「振ったサイコロは理想的なサイコロである」というのは、「仮説」だし、「振ったサイコロは６つの目のどれも等確率で出る」と言うのも仮説です。「振ったサイコロは 1,2,3,4,5,6 の目を持ち、それぞれの目が出る確率は $p_i = \frac{1}{6}; i = 1, 2, ..., 6$ である」と変数を持ちだして、確率的に起きる現象であることを説明してあるとモデルらしさが強いという感じでしょうか。

2.3　尤度関数と対数尤度関数

「サイコロを 3 回振ったら、4,2,4 が出た」と言う事象は「サイコロを 3 回振ったら、1,2,3,4,5,6 の目がそれぞれ、$\mathbf{x} = (0, 1, 0, 2, 0, 0)$ 回ずつ出た」ともいいかえられます。「6 つのサイコロの目の出る確率が $\mathbf{p} = (p_i); \sum_{i=1}^{6} p_i = 1; i = 1, 2, ..., 6$」であるというモデルの尤度は

$$L(\mathbf{p}|\mathbf{x}) = \binom{\sum_{i=1}^{6} x_i}{x_1, x_2, ..., x_6} \prod_{i=1}^{6} p_i^{x_i}$$

となる。この関数を尤度関数と言います。$p_i^{x_i}$ のようなべき乗があると計算が面倒臭いので、対数を取ってやる。対数を取ったものを対数尤度関数と言います。

$$LL(\mathbf{p}|\mathbf{x}) = \ln L(\mathbf{p}|\mathbf{x}) = \ln \binom{+\sum_{i=1}^{6} x_i}{x_1, x_2, ..., x_6} \sum_{i=1}^{6} x_i \ln p_i$$

2.4　変数の数と自由度

ここで、サイコロの目の出る確率 $\mathbf{p}$ について、2 つのモデルを考えます。一つ目のモデルは、「理想的なサイコロであるモデル」で、$\mathbf{p^1} = (\frac{1}{6}, \frac{1}{6}, ..., \frac{1}{6})$ です。二つ目のモデルは、なんでもいいから適当に決めてよいモデル $\mathbf{p^2} = (p_i)$ です。一つ目のモデルは 6 つの確率の値 p_i のどれも、自由に決めることができませんでした。二つ目のモデルは、$\mathbf{p^2}$ の 6 つの確率の値のうち $p_1, ...p_5$ までを自由に (ただし $p_1, ..., p_5 \geq 0$ で、$p_6 = 1 - \sum_{i=1}^{5} p_i \geq 0$ という制約はありますが、その範囲では自由に) 決めることができました。かたや、まったく自由がなく、かたや、5 つが自由でした。この自由に値を選べる変数の数の差を自由度と呼びます。二つ目のモデルは一つ目のモデルに比べて、自由度が 5 だけ大きいのです。この自由度は後に出てくる尤度比検定のところで使います。

2.5　最尤推定値

たとえば $\mathbf{p^2} = (p_1, p_2, ..., p_6) = (0, \frac{1}{3}, 0, \frac{2}{3}, 0, 0)$ とするのも自由です。これは、二つ目のモデルの尤度が最も大きくなるように選びました。このように尤度を最大にするモデル変数の値を最尤推定値と呼びます。最尤推定値を尤度関数の微分で出すことにすていは 11 章を参照。

2.6　尤度比はカイ二乗分布に照らして検定することができる

２つのモデルがあって、両者の違いは自由な変数の個数であるとき、変数が自由なモデルの方が尤度が大きく、従って対数尤度も大きい。２つのモデルの対数尤度の差 (２つのモデルの尤度の比の対数を取ったものなので、対数尤度比と言う) の２倍の値が有用になる。

$$2\ln\frac{L(\mathbf{p^2}|\mathbf{x})}{L(\mathbf{p^1}|\mathbf{x})} = 2(LL(\mathbf{p^2}|\mathbf{x}) - LL(\mathbf{p^1}|\mathbf{x}))$$

この統計量は帰無仮説 (サイコロが理想的であるという仮説) が成り立っているときには、自由な変数の数を自由度としたカイ二乗分布に従うことが知られている。従って、この統計量を用いて帰無仮説を棄却すべきかどうかの検定 (棄却検定) が可能である。R で検定してみる。

```
> p1 <- rep(1/6,6)
> p2 <- c(0,1/3,0,2/3,0,0)
> x <- c(0,1,0,2,0,0)
> LL1 <- lgamma(sum(x)+1)-sum(lgamma(x+1)) + sum(x*log(p1))
> LL2 <- lgamma(sum(x)+1)-sum(lgamma(x+1)) + sum(x[which(x!=0)]*log(p2[which(x!=0)]))
> S <- 2 * (LL2-LL1)
> df <- 5
> pchisq(S,df,lower.tail=FALSE)

[1] 0.2257869
```

帰無仮説が成り立っているときに尤度比検定を繰り返すと、p 値が一様分布する様子は、12 章を参照。

2.7　どんなときに用いるか

尤度比検定は「尤度比」という言葉が入っているけれども、多くの DNA 鑑定の場合のように、２つの仮説の尤度の比を比べているときには用いません。なぜなら、DNA 鑑定での２つの尤度は、それぞれ「決め打ちにした仮説」であるのに対して、尤度比検定の方の尤度は、かたや、何かしらモデル変数の自由を奪って制約したもの、かたや、モデル変数を自由に動かしたものです。従って、対立仮説の方の尤度は、自由に変数の値を動かせ

る世界の中で、最大の尤度になっています。DNA 鑑定の尤度比検定では「もう 1 つの仮説」の尤度が、何かしらの自由な世界の頂点であるというわけではない点が違います。実際、尤度比検定で統計量をカイ二乗統計量に照らして評価するのは、変数を自由に動かす世界に連続的充満している仮説群があることを前提にしたいるから可能であるのです。では、この尤度比検定は DNA 鑑定では用いないのか、というと、そうではありません。実験データの評価をするときに変数を用いたモデルを立てたり、ベイジアンネットワークでモデル変数を用いたりするときには、立てたモデルが意味のあるものであるのかどうかの評価に尤度比検定の考え方を使います。サイコロの例でもみたとおり、変数を自由にすると尤度は大きくなりますが、「変数を自由にしたのなら、それに見合うだけ十分に大きくならないのなら、その尤度の増加は意味のないものかもしれない。意味のありなしを尤度比検定の考え方で評価しよう」という使い方です。

第 3 章

家系図での DNA 鑑定用尤度計算法について 2011 年 5 月

3.1　はじめに

2011 年 3 月 11 日、東日本大震災とそれに引き続く大津波があり、多数の方が亡くなられました。私ごとながら、出生地である宮城県多賀城市も大きな被害を受けました。さて、この大震災により、多人数の行方不明者と多数のご遺体が県境を越える形で身元鑑定を必要とする事態が発生しました。DNA 鑑定はその強力なツールの一つとなりました。実務上、いくつかの課題があると法医学講座より情報提供を受け、それにそって取り組んでみました。まずは、血縁者からの DNA 提供が得られている状況での行方不明者のジェノタイプ推定とそれに基づく尤度計算です。また、複数の行方不明者と複数のご遺体との多対多マッチング問題がありました。また、ジェノタイプ情報はそれ以外の情報 (ご遺体の発見場所・身体的特徴) などと併せて解釈されるべきであることも問題となりました。これらについて、何回かに分けて法数学勉強会でも取り上げました。本章はそのうち、血縁情報を用いた DNA 鑑定用の尤度計算法に関してになります。

3.2　ジェノタイプ確率の計算の基礎

勉強会で使用したスライドはこちらです。

ある家系図があって、その家系がある集団に属しているとき、家系図上のメンバーのジェノタイプを活用して、ジェノタイプ不明なメンバーのジェノタイプの確率を計算する

とは、どのような手順で行われているのでしょうか？一つのやり方を示します。

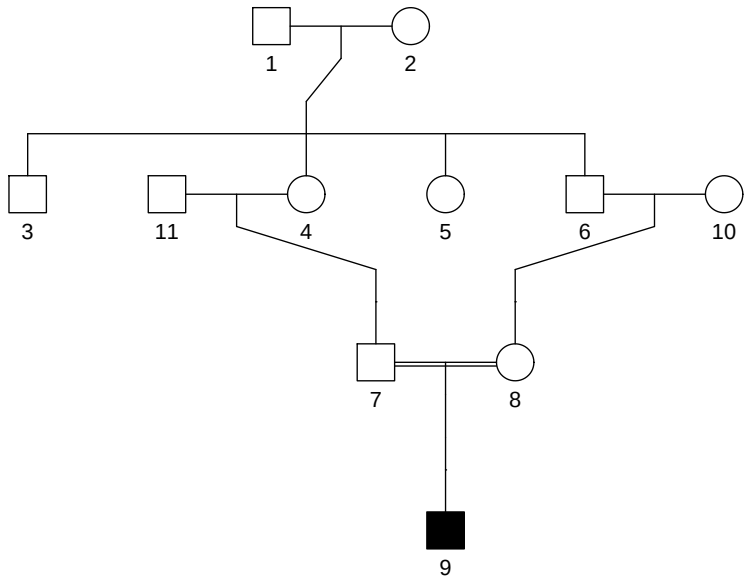
- 家系図を描く
  - ジェノタイプを提供しているすべての個人をプロットする
  - 親子関係をつなぐ
  - 片親しかいない場合には、もう一人の片親を仮に描いてつなぐ
  - 親子以外の血縁関係も途中にはいるべき血縁者を仮に描いておく
- 親子関係をすべて列挙する
  - 親子関係は両親と子のトリオ
  - ジェノタイプのない個人もすべて「仮」に描いてあるので、すべての人は次の二通りに分かれる
    - 親が２人とも描かれている
    - 親がどちらも描かれていない
  - 親がどちらも描かれてない人は、「集団から生まれた」とみなす。
- 個人のジェノタイプを確率的に決める
  - ジェノタイプが提供されている人の場合は、１つのジェノタイプの確率が 1
  - 「集団から生まれた人」のジェノタイプは集団のジェノタイプの頻度に応じて確率的に決まる
  - それ以外の人 (両親がともに家系図に描かれているが、本人のジェノタイプは提供されていない人) のジェノタイプは、確率的に計算する

問題は「それ以外の人 (両親がともに家系図に描かれているが、本人のジェノタイプは提供されていない人) のジェノタイプは、確率的に計算する」を実際にどうするか。それについては細かい話になるので、13 章を参照。ただし、ジェノタイプの確率的推定方法に関して、家系図は２種類に分かれることは有用な知識と思いますので、コメントしておきます。

家系内に近親婚がある場合、これは、家系図をグラフで描いたときに閉じた輪がある場合です。図では７と８の間の関係がいとこ婚です。『９→７→４→（１，２）→６→８→９』という輪ができていて、これが家系内の近親婚に対する輪です。ジェノタイプの確率計算においては、このような『輪』がない場合には、簡単ですが、『輪』があると面倒臭くなります。これはグラフ (家系図は個人 (や個人が持つ染色体) を点とし、伝達関係を辺としたグラフ) を扱う理論であるグラフ理論において、サイクル (ぐるりとめぐる『輪』) があるかないかで、アルゴリズムが大きく変わることに対応しています。グラフ理論は DNA 鑑定領域であれば、ベイジアン・ネットワークもグラフです。ベイジアン・ネットワークで

もグラフのサイクルの有無は問題にされます。

```
> library(kinship2)
> test1 <- data.frame(id  =c(1,2,3,4,5,6,7,8,9,10,11), mom =c(0, 0, 2, 2,2, 2, 4, 10,8,0,0), dad =c(0, 0, 1, 1, 1, 1, 11,6,7,0,0)
> affected<-rep(0,11)
> affected[9] <- 1
> ptemp<-pedigree(id=test1$id,dadid=test1$dad,momid=test1$mom,sex=test1$sex,affected=affected)
> plot(ptemp)
```



第 4 章

多人数一括 DNA プロファイリング手法 2011 年 9 月

4.1　はじめに

本章は、前章に引き続き大震災関連課題を取り扱います。多人数の行方不明者と多人数のご遺体とのマッチングに関する話です。

4.2　１人の DNA 鑑定

勉強会で使用したスライドはこちらです。

ある行方不明者 m のジェノタイプがわかっており、ある遺体 b のジェノタイプがわかったとき、本人確認は、それが一致するかどうかを基にします。通常の刑事上の DNA 鑑定とだいたい同じ枠組みです。m のジェノタイプがわかっておらず、その血縁者のジェノタイプが得られたとき、m のジェノタイプは一意に決まらず、確率的に決まりますが、それは計算できます。もし、m が１人で、b が１体であったなら、

- m は b である
- m 以外の誰かが b であって、m は b 以外の誰かである

という２つを問題にすることになります。

4.3　複数人の場合

複数の人 $M = \{m_1, m_2, ..., m_{Nm}\}$ が行方不明や犠牲になり、複数のご遺体 $B = \{b_1, b_2, ..., b_{Nb}\}$ が見つかった場合というのもある。今、行方不明者数とご遺体数が同数であるときを考えます。考えるべき仮説は

- $m_1 = b_1, m_2 = b_2, m_3 = b_3, ..., m_{Nm} = b_{Nm}$
- $m_1 = b_2, m_2 = b_1, m_3 = b_3, ..., m_{Nm} = b_{Nm}$
- ...

となりますが、この仮説の数は

$$N_m! = N_m \times (N_m - 1) \times (N_m - 2) \times ... \times 2 \times 1$$

となります。それぞれの仮説の下で観察データ (この場合は、M のジェノタイプ (血縁者試料からの推定ジェノタイプを含む) と B のジェノタイプのこと) を観察する尤度が計算できます。原理としては、これを比較すればよいわけです。

4.4　割り付け問題

複数人について $N!$ 通りの仮説を考えてそれぞれの尤度を計算して、尤度が最大のものを選び出す、というのは、言うのは簡単ですが、N の数が少し多くなると、$N!$ がものすごく大きくなるので、非現実的です。

```
> N <- 1:20
> N
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
> factorial(N)
 [1] 1.000000e+00 2.000000e+00 6.000000e+00 2.400000e+01 1.200000e+02
 [6] 7.200000e+02 5.040000e+03 4.032000e+04 3.628800e+05 3.628800e+06
[11] 3.991680e+07 4.790016e+08 6.227021e+09 8.717829e+10 1.307674e+12
[16] 2.092279e+13 3.556874e+14 6.402374e+15 1.216451e+17 2.432902e+18
```

$N = 5$ で 120 通り、$N = 10$ で 300 万通りを超え、$N = 12$ で約 5 億通りです。しかしながら、何人かが集まって、その人たちをうまく割り当てたい、という希望は、DNA 鑑定に限らず、日常茶飯なことなので、この手の課題は「割り付け問題」と呼ばれ、比較的うまく対処する方法があります。たとえば、$N!$ 通りのすべてについてしらみつぶしに調べることは無理だけれど、最もよい割り付け方を見つけ出す、というアルゴリズムはあって、コンピュータを使えばわかります。14 章に簡単に計算できる様子を示しました。

4.5 最適割り付けが答えなのか

最もよい割り付け (最適割り付け) を $N!$ 通りから見つけ出せたら、それが「正解」なのでしょうか？最適割り付けは「尤度」が「最も高い割り付け方である」ということしか言っておらず、「他の割り付け方はあり得ない」と言っているわけではありません。もしも、2 番目に高い尤度が、最大尤度よりも明らかに小さいことがわかれば、安心して最大尤度をもたらす割り付けを採用することができます。最大尤度をもたらす割り付けを見つけるアルゴリズムはあると書きましたが、上から $k = 1, 2, ...$ 番目までの尤度をもたらす割り付けまでを見出すアルゴリズムもあります。もちろん k がそれほど大きくない場合です。これができるのであれば、$k = 1, 2$ を調べ、その 2 つの尤度に大差があれば、選ぶべき割り付けは決まります。

4.6 最適割り付けを選べない場合

もし $k = 1, 2$ の割り付けを見つけ、$k = 1$ を選ぶことが適切であると決まらなかったらどうしたらよいでしょうか？決まらない、と言うのは、$k = 1, 2$ の尤度がそれほど違わない場合で、$k = 1$ の方が尤もらしいけれど、$k = 2$ である可能性を捨てきれない、という場合です。

4.7 DNA 鑑定でほしい答え

すべての行方不明者とすべてのご遺体との対応が一発でわかれば、とても素晴らしいですが、行方不明者を探している人の立場からすれば、

- 私の探している人はこのご遺体なのか、そうでないのか
- 私の探している人はこのご遺体であるかもしれない、と言うことだが、それ以外のご遺体である可能性はないのか

というように、もっぱら、自分の探している人について興味があるだけで、それ以外の行方不明者とご遺体とのマッチング関係については興味がほとんどないでしょう。また、この質問に答えることが、ご遺体を遺族が引き取ることができるかの決断のためになすべきことであることも多いでしょう。では、この個別の希望に対して、割り付け問題はどのようにアプローチすることができるでしょうか？今 N 人の行方不明者とご遺体があるときに m_i という行方不明者についての質問に答えたいとします。$m_i = b_1, m_i = b_2, ..., m_i = b_N$ という N 通りの仮説があります。割り付け問題では $N!$ 通りの仮説があったのに比べると、ずいぶん少ない数です。ただし、$m_i = b_j$ という仮説は、実際には m_i 以外の $N-1$ 人の行方不明者と b_j 以外の N_1 体のご遺体とがどのように対応づけられるかのすべてが合算されています。$N-1$ 人と $N-1$ 体の総当たり割り付け問題が封じ込められているわけです。ここには $(N-1)!$ 通りの割り付け方法がありますから、この確率・尤度をすべて計算することにしたのでは、場合の数が大きすぎる問題の解決にはなりません。そうではなくて $(N-1)!$ 通りの確率・尤度を個別に計算して足し合わせる代わりに、すべての確率・尤度の和をいっぺんに計算することができれば、場合の数問題に対応したことになります。そのような方法は規模がそれほど大きくなければないわけではないようです。そんなアプローチが適切であるかもしれません。

第 5 章

DNA 鑑定とそれ以外の情報の組合せのための基礎 2011 年 11 月

5.1 はじめに

本章も大震災関連話題です。本章では DNA 情報とその他の情報の組合せ方に関する考え方を扱います。ただし、このテーマは震災のみならず、広く DNA 鑑定一般につながる内容です。

第 6 章

犯人である確率を正確に計算する～ちびまる子ちゃん事件～ 2012 年 3 月

6.1　はじめに

棄却検定では正確確率検定と呼ばれる方法があります。本章では、得られた情報から分割表を作成し、正確確率を計算することで、ある容疑者が犯人である確率を正確に計算することについて考えます。

6.2　ちびまる子ちゃん事件

勉強会で使用したスライドはこちらです。

ある小学校では、すべての男子生徒３００人に「消しゴム屋」さんから１個ずつ消しゴムが配られた。
「ナルト」柄が５０個、「ワンピース」柄が１００個、「ドラえもん」柄が１５０個
ある日、音楽室で、１個の消しゴムの落し物が発見された。
「ナルト」柄だった（５０個/３００個）
その消しゴムには「ほなみたまちゃん LOVE」と書かれていた！
さあ、この消しゴムは誰のものか？
女子児童たちの捜査が始まった。

捜査員の一人であるみぎわさんが消しゴム屋に迫った。
「消しゴム屋さん、『ナルト』柄の消しゴムを配った男子のリストを出しなさいよー」と。
リストが得られれば、『ナルト』所有者 50 名に絞られるからだ。そのリストに載っている 50 名は、1/50 の確率で「たまちゃん LOVE」であるとわかるはずだ。
しかし、その情報は得られない。「あいにく、個人情報保護法の規定により、リストをお出しすることはできません」と、消しゴム屋さん。ただし、「特定の児童さんを指名していただけば、その児童さんにどのタイプを渡したかはお答えできます」と。

「たまちゃん LOVE」の『犯人』って誰だと思う？
「やっぱり、同じクラスの男子？」
「クラブが一緒？」
「委員会つながりかも？」
「以前に同じクラスだった・・・とか」
「帰り道が一緒だからだったりして」

と、女子全体の詮索が続く。
それを聞いた、まるちゃん母、
「事前確率なんて考えるのはやめなさい、下手な考え、休むに似たりって言うじゃない」

そうとばかりは言えないのだが…
そんなところへ、音楽の先生から耳より情報が得られた。
「音楽室はいつも鍵がかかっていて、鍵は私しか持っていないから、その日、音楽室に出入りできた児童は、その日に音楽の授業があったクラスの子、だけよ」と。

「その日、音楽の授業があったのは、３クラス、１２０人のうち、男子は、６０人」
「３００人のうち、６０人が候補者だ」

話を整理しよう。

- 「その日、音楽の授業があったのは、３クラス、１２０人のうち、男子は、６０人」
- 「音楽の授業は、３００人のうち、６０人」
- 「『ナルト』柄は、３００人のうち、５０人」
- 「じゃあ、音楽のあったクラスで、『ナルト』消しゴムを持っていたのは何人？」

ここから得られのは
「６０ｘ５０/３００　＝１０人　くらいね！」

　さて、消しゴム屋さんからの情報も使うとしよう。
「特定の児童さんを指名していただけば、その児童さんにどのタイプを渡したかはお答えできます」と。
みぎわさんは
『絶対、花輪君のを教えてもらうわ！誰も文句ないわよね！』
と方針を決定。

　それを受けて、消しゴム屋さんいわく
「花輪君ですね、花輪君は、『ナルト』柄でした」

　みぎわ：「!弩?恐狂!!逆散魔*警鱗!鬱 + 閉静無.. 止....」

6.3　ちびまるこちゃんちの評定

6.3.1　まるちゃん母の意見

　「花輪君以外の男子 (５９人) が、『ナルト』柄である確率は $50/300 = 1/6 \to 49/299$」
「そのうちの誰かが、落とし主だとすると、それが『ナルト』柄である確率は $49/299$」
「花輪君が落としたのなら、それが『ナルト』柄である確率は　1」
「誰かが落としたのなら、それが『ナルト』柄である確率は $49/299$」
「尤度比は $1/(49/299) = 299/49 = 6.10$」
「花輪君が落としたと考える方がよさそうじゃない？」

　みぎわ：「叫怒天髪射」

6.3.2　まるちゃん父の意見

　「候補の男子が 59 人もいるのに、いきなり、『 6.10 倍』っていうのは、おかしいんじゃねーか？」

「花輪君が落としたのなら、それが『ナルト』柄である確率は　1」
「誰か１が落としたのなら、それが『ナルト』柄である確率は 49/299」
「誰か２が落としたのなら、それが『ナルト』柄である確率は 49/299 」
…
「誰か 59 が落としたのなら、それが『ナルト』柄である確率は 49/299 」
「花輪君 vs. その他 59 人 は 1: 59x49/299 = 0.103 」

みぎわ :「莞喜爾天静昇」

6.3.3　まるちゃん祖母の意見

「音楽の授業は、３００人のうち、６０人」
「『ナルト』柄は、３００人のうち、５０人」
「じゃあ、音楽のあったクラスで、『ナルト』消しゴムを持っていたのは何人？
６０x５０/３００　＝１０人　くらいね！

じゃったろう
「本当のところは、音楽授業の６０人中、何人が『ナルト』柄じゃったんかいのー」
１人、２人、…、５０人の場合が全部ありえるはずじゃなぁ
１人だけが『ナルト』柄で、それが花輪君なら、「たまちゃん LOVE」は花輪君で確定じゃ
２人が『ナルト』柄なら、花輪君かもしれないし、もう１人かもしれなくて、確率は 1/2
…
k 人が『ナルト』柄で、花輪君がそのうちの１人なら、確率は 1/k
これを計算するとどうなるかの？

6.4　まるちゃん祖母の計算

『ナルト』柄の確率は

$$p = \frac{50}{300} = \frac{1}{6}$$

音楽室に出入りした 50 人中、k 人が『ナルト』柄である確率は

$$pr(k) = \frac{50!250!60!240!}{300!k!(50-k)!(60-k)!(190+k)!}$$

こんな 2 × 2 表で考えれば計算式はわかりやすい。

	Naruto	Non-Naruto	行和
Ongaku-shitsu	k	$50-k$	50
Not-Ongaku-shitsu	$60-k$	$190+k$	250
列和	60	240	300

k の大きさに比例して『ナルト』柄が落ちやすくなるから

$$Pr(k) = k \times pr(k) = k \times \frac{50!250!60!240!}{300!k!(50-k)!(60-k)!(190+k)!}$$

k 人の誰もが「たまちゃん LOVE」と書く確率は同じであれば

$$Hanawa(k) = \frac{1}{k} \times k \times pr(k) = \frac{50!250!60!240!}{300!k!(50-k)!(60-k)!(190+k)!}$$

$k = 0$ の場合は『ナルト』柄消しゴムが見つかっているから、もうありえないので、すべての場合の和は

$$\sum_{i=1}^{50} Pr(k) = \sum_{i=1}^{50} k \times \frac{50!250!60!240!}{300!k!(50-k)!(60-k)!(190+k)!}$$

花輪君が「たまちゃん LOVE」であるのは

$$\sum_{i=1}^{50} Hanawa(k) = \sum_{i=1}^{50} \frac{50!250!60!240!}{300!k!(50-k)!(60-k)!(190+k)!}$$

この方法では

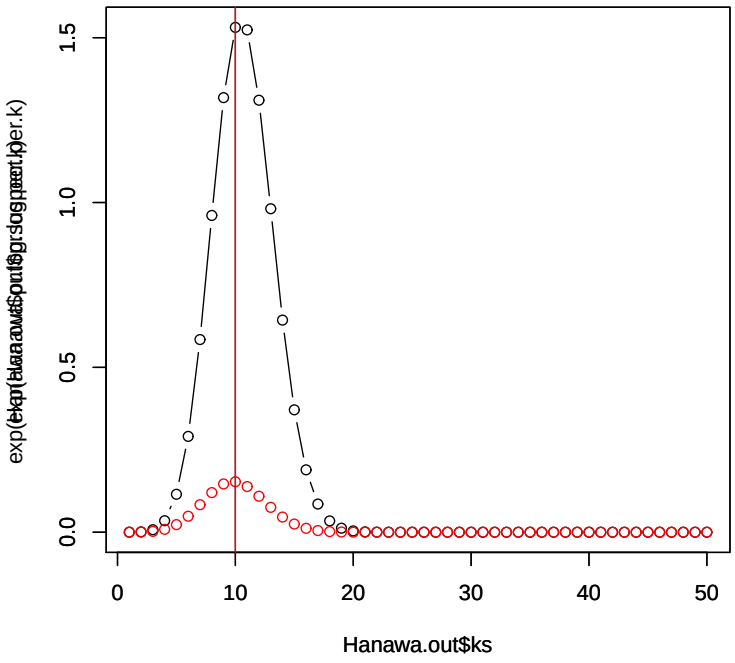
花輪君が「たまちゃん LOVE」である尤度は，0.100

花輪君が「たまちゃん LOVE」である尤度と、他の誰かが「たまちゃん LOVE」である尤度の比は、0.111

```
> mySuspect<-function(p,N){
+       ks<-0:N
+       if(p>=1){
+               ks<-N
+               Prk.log<-lgamma(N+1)-lgamma(ks+1)-lgamma(N-ks+1)+ks*log(p)+log(ks)
+               PrHanawa.log<-Prk.log-log(ks)
```

```
+                 LikeHanawa<-0
+                 LikeAll<-0
+                 Pr.Hanawa<-0
+                 LR.Hanawa<-0
+         }else{
+                 Prk.log<-lgamma(N+1)-lgamma(ks+1)-lgamma(N-ks+1)+ks*log(p)+(N-ks)*log(1
+                 PrHanawa.log<-Prk.log-log(ks)
+                 LikeHanawa<-sum(exp(PrHanawa.log[2:length(ks)]))
+                 LikeAll<-sum(exp(Prk.log)[2:length(ks)])
+                 Pr.Hanawa<-LikeHanawa/LikeAll
+                 LR.Hanawa<-LikeHanawa/(LikeAll-LikeHanawa)
+         }
+
+         return(list(pr.log.per.k=Prk.log,pr.log.suspect.per.k=PrHanawa.log,like.all=LikeAll,like.
+ }
> mySuspect.Exact<-function(m1,m2){
+         N<-sum(m1)
+         ks<-0:min(m1[1],m2[1])
+         M<-cbind(ks,m1[1]-ks,m2[1]-ks,m1[2]-m2[1]+ks)
+         minM<-apply(M,1,min)
+         selected<-which(minM>=0)
+         M<-M[selected,]
+         selected2<-which(M[,1]>0)
+         M<-M[selected2,]
+         tmp<-sum(lgamma(m1+1),lgamma(m2+1))-lgamma(N+1)
+         Prk.log<--apply(lgamma(M+1),1,sum)+tmp+log(M[,1])
+         PrHanawa.log<-Prk.log-log(M[,1])
+         LikeHanawa<-sum(exp(PrHanawa.log))
+         LikeAll<-sum(exp(Prk.log))
+         Pr.Hanawa<-LikeHanawa/LikeAll
+         LR.Hanawa<-LikeHanawa/(LikeAll-LikeHanawa)
+
+
```

```
+         return(list(pr.log.per.k=Prk.log,pr.log.suspect.per.k=PrHanawa.log,like.all=LikeAll,like.suspect=LikeHanawa,pr.sus
+ }
> p<-50/300
> N<-60
> m1<-c(50,250)
> m2<-c(60,240)
> Hanawa.out<-mySuspect.Exact(m1,m2)
> ylim=c(0,max(exp(Hanawa.out$pr.log.per.k)))
> plot(Hanawa.out$ks,exp(Hanawa.out$pr.log.per.k),type="b",ylim=ylim)
> par(new=TRUE)
> plot(Hanawa.out$ks,exp(Hanawa.out$pr.log.suspect.per.k),type="b",col=2,ylim=ylim)
> abline(v=which(Hanawa.out$pr.log.per.k==max(Hanawa.out$pr.log.per.k)))
> abline(v=which(Hanawa.out$pr.log.suspect.per.k==max(Hanawa.out$pr.log.suspect.per.k)),col=2)
```



6.5　終わりに

場合によっては作成される分割表が複雑になります。そのように複雑になった分割表では、周辺度数による分割表制約がベイズ流の事後確率としての性質を強くすることもわかりました。

第 7 章

事前確率と共役事前分布 2012 年 9 月

7.1　はじめに

これまではジェノタイプ情報やその他の情報を用いてある仮説がどれくらい尤もらしいかを考えてきました。実際のところ、この尤もらしさの評価が提供しているのは尤度比であって、それは、事前確率を事後確率に変化させる要素であって、事前確率が不明なときには、利用の意味が大きく損なわれます。従って、DNA 鑑定においてジェノタイプ情報を的確に活用するためにはジェノタイプがもたらす尤度比のみではなく、その一歩手前の事前確率に関して、どのようなことを統計学・数学が提供できるかを考えることは重要です。本章はこの点に関する話です。

7.2　事前確率は心の中にある

事前確率は情報がない状態もしくは、情報が限られている状態で思い描く、仮説の確率のことです。例でそのことを確認します。

7.2.1　最高気温の予想

ある夏の日の最高気温が何度になるかを予想するとします。
最高気温ですから、20-40 度くらいの何かしらの値になるでしょう。今、これ以外に情報がないとき、「20-40 度」くらい、という気持ちが事前確率です。
ただし、20 度の確率はどれくらいで、30 度の確率はどれくらいで 35 度の確率はどれく

らいで 40 度の確率はどれくらいなのか、32.3 度の確率はどうなのか、と連続した実数値に対して、「思い入れ」の強弱のカーブが描けるはずです。可能性の話をしますから、絶対零度から無限大までという取りうる温度の数値のすべてに対して、「思い入れ」の強弱をつけ、それを足し合わせ (積分) したときに 1 になるように、スケールを調整したもの、それが事前分布です。

情報

朝、 6 時半、外から元気に体操する声が聞こえたとします。その情報から、雨が降っていないようだし、『朝、晴れていた夏の日の最高気温』になりそうだと考えて、少し情報修正するかもしれません。このように情報修正したとすれば、「体操」という情報を使って、事前確率の分布を事後確率の分布に変えています。このようにして得た事後確率の分布も、次に「クマゼミがいっせいに『シャワシャワ』と鳴く声が聞こえた」という情報を受けて、さらに上方修正されるかもしれません。ですから、事後確率はさらなる情報に関して事前確率として働きます。
このように情報に応じて変化する事前確率の分布ですが、どのような分布を描くかは、予想する人の経験等にも左右されますから、個人個人で異なります。

7.2.2　事前確率の分布に客観性を持たせる

心の中にある事前確率分布、と書いてきましたが、予想するべき場所と日付がわかれば、過去の気象記録を活用することは、いい手でしょう。天気などについてまったくわからなければ、天気などによらず、過去の最高気温分布をそのまま使うのが良さそうです。もし、予想したい日が朝から晩まで晴れと思われているなら、過去にあったそんな日の最高気温だけを使えば良いでしょう。

7.2.3　父親である事前確率

孤島にサルの群れが暮らしていて、あるメス猿から子ザルが生まれたとします。この子ザルの父親はどのオス猿なのか、ということを知りたいとします。サル山にボス猿がいれば、それが父親である確率は相当高いことでしょう。どれくらい高いかは、「心の中」にあるかもしれないですし、何かしらの調査の結果に基づいて、ある程度の絞り込みができるかもしれません。
もし、この子ザルがちょうど「ボス猿不在の時代」の生まれだとします。どのように考えればよいでしょうか。相変わらず「心の中」に分布はありますが、もし、母猿とオス猿す

べての接触時間に関する情報が得られるのであれば、それを客観的に活用して、各オス猿について「父親である確率」の高低を調整することができるでしょう。

7.2.4 シンデレラの靴

シンデレラが落として行ったガラスの靴を頼りに、王子がシンデレラを探すのも、靴を用いて、女性の事前分布を変えて、それらしい女性に絞り込むためです。同じようなことはそこここでやられています。ただし、それを行う時に、「シンデレラではありえない」「シンデレラであるかもしれない」の２つにはっきりと分けることができるか、最高気温予想のように、どの値がどれくらいありそうかを相対的に重みづけするかでは、少し考え方が違うでしょう。DNA 鑑定の場合には、多くの場合、はっきり２分するというよりは、色々な仮説の事前確率が高かったり低かったり色々だけれど、0 であることはない、としておくことが重要です。しかしながら、DNA 鑑定を用いて、判決という決断を下さなければならないとき、「決断」は 0 か 1 かの選択をすることですから、どんなに頑張っても、ある程度の曖昧さを切り捨てる「痛み」が生じることはやむをえないと考えることが適当です。

7.3 共役事前分布

事前分布のことを扱ったので、共役事前分布にも触れておきます。共役事前分布はベイジアン・ネットワークなどでよく使うものです。例を出します。画鋲を投げたときに、針が上を向いて止まる場合と針が下を向いて止まる場合とがあります。今、画鋲が針を上に向けて止まる確率 p、下に向けて止まる確率 $1-p$ について知りたいとします。

画鋲を投げる実験を開始する前に、p の予測を「心」に聞いてみます。それを実験前の予測分布として p_0 とします。p_0 は 0 から 1 の値を取る分布であるはずです。どんな分布を思い描いてもよいのですが、「ベータ分布」というのを思い描くと便利ですよ、というのが、「共役事前分布」とは何か、という話とつながります。

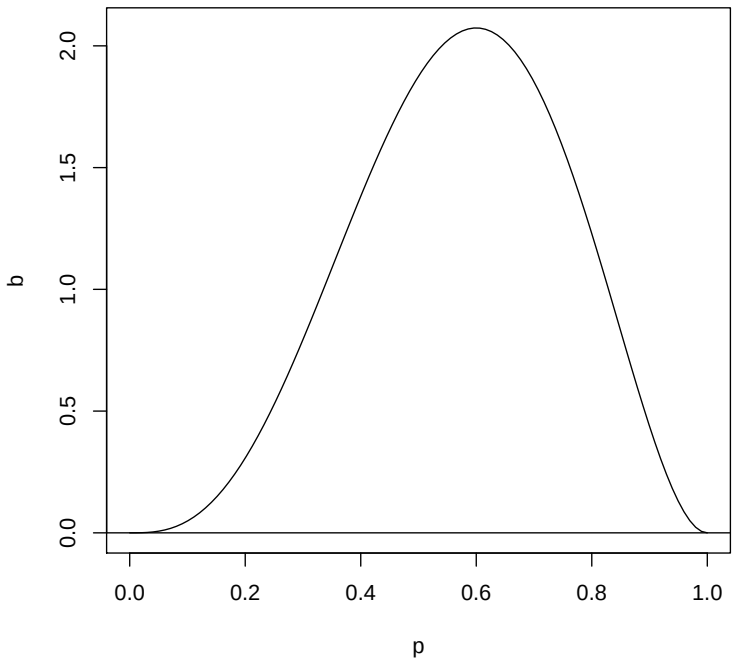
7.3.1 これも画鋲の表と裏〜ベータ分布〜

画鋲の裏 (針が下を向いている状態) が出る確率が 0 である確率はないだろうし、1 である確率もないだろう。0.5 ということもなさそうだ。0.5 よりは大きいところにピークを持たせて、$p = 0, 1$ で 0 になるような分布が描きたい。

ベータ分布と呼ばれる分布は次の図のような形をしており、0 から 1 の間で積分すると 1

になるという性質がある。形と言い、積分すると 1 になる性質といい、事前分布としてとてもよい。

```
> p <- seq(from=0,to=1,length=100)
> b <- dbeta(p,4,3)
> plot(p,b,type="l")
> abline(h=0)
```



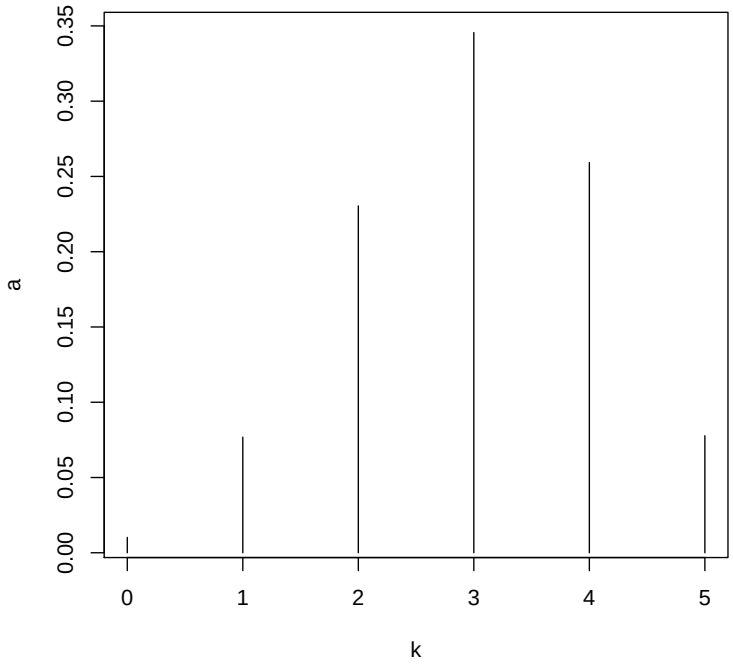
7.3.2　これも画鋲の表と裏～二項分布～

実験を開始しました。n 回投げたら、k 回の裏、$n-k$ 回の表が出るでしょう。裏の出る確率を p としたらこの確率は

$$\binom{n}{k} p^k (1-p)^{n-k}$$

となります。これが二項分布です。$n = 5, k = 0, 1, 2, 3, 4, 5, p = 0.6$ の場合のそれを見てみます。

```
> n <- 5
> k <- 0:5
> p <- 0.6
> a <- dbinom(k,n,p)
> plot(k,a,type="h")
```



7.3.3 画鋲の表と裏に関する２つの分布の関係

p の値を決めると二項分布が描けた。p を決めるというのは、仮説を決めたことになり、二項分布を描いたのは、その仮説の下での生起確率を計算したことになります。生起確率があれば、尤度があることは冒頭の章でも出てきた内容ですから、それについて考えてみます。尤度とは、ある仮説が、ある観察結果をもたらす確率のことですが、その確率を仮

説の側からみたものです。ある特定の仮説ではなく、すべての仮説について、ある特定の観察結果をもたらす確率がわかるとよいです。画鋲の表裏の場合の仮説とは p が 0 から 1 のどれかです。p は連続値ですから、$0-1$ の範囲で連続関数のはずです。実は、このような関数がベータ分布です。

実際、その計算式は

$$\binom{n}{k} p^k (1-p)^{n-k}$$

と同様です。ただし、p について $0-1$ で積分をして 1 にならないといけませんから、

$$\frac{p^k (1-p)^{n-k}}{\int_0^1 t^k (1-t)^{n-k} dt}$$

となります。

7.3.4　何が便利なのか

今、画鋲を実際に n 回投げて k 回の裏 $n-k$ 回の表が出た、という情報があれば、p の尤度はベータ関数で計算できるから、それをそっくりそのまま使えば良いです。

しかしながら、昔、画鋲を投げたとき、裏が K 回、表が $N-K$ 回くらいだったな、と言うような、おぼろげな記憶があったとしたら、どうしたらよいでしょうか？ベータ分布で実際に投げた n, k を使って

$$\frac{p^k (1-p)^{n-k}}{\int_0^1 t^k (1-t)^{n-k} dt}$$

とする代わりに、実験前の事前分布を

$$\frac{p^K (1-p)^{N-K}}{\int_0^1 t^K (1-t)^{N-K} dt}$$

と想定してやれば、その後、実際に実験をした後に信じるべきなのは

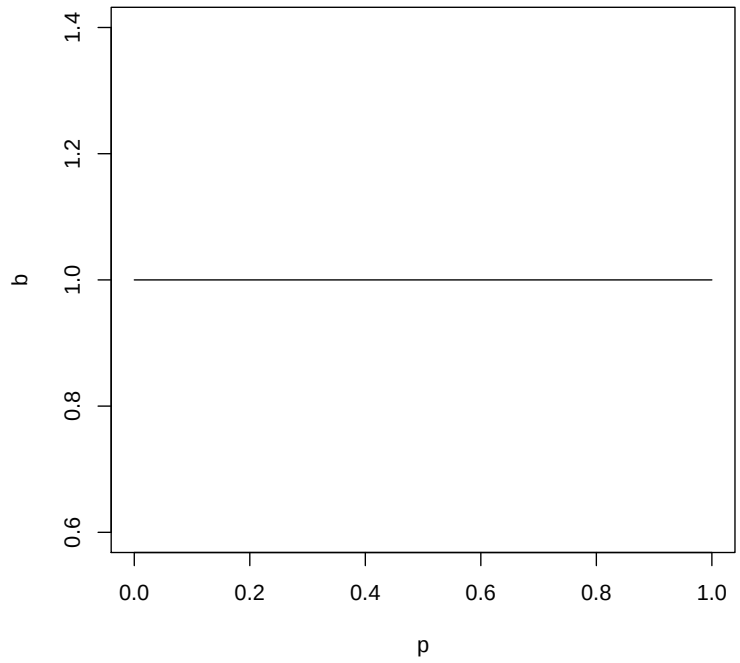
$$\frac{p^{K+k} (1-p)^{(N+n)-(K+k)}}{\int_0^1 t^{K+k} (1-t)^{(N+n)-(K+k)} dt}$$

と簡単に指数を変えるだけで済みます。また、実験結果だけを用いて、p の値を推定するよりも、実験する回数を減らしてもよくなるかもしれません。

逆に言うと、実験結果だけを用いて p の分布を推定する、というのは、事前確率として、おぼろげな記憶がない、つまり $N=0, K=0$ という状態からスタートして予想する、と

いうことでもあります。たしかに $N = 0, K = 0$ のベータ分布は次に示すように $0 - 1$ のどれも等確率になった分布のことです。

```
> p <- seq(from=0,to=1,length=100)
> b <- dbeta(p,1,1)
> plot(p,b,type="l")
```



7.3.5 共役事前分布

ベータ分布は二項分布の共役事前分布 (二項分布を観測の分布としたとき、仮説の側の分布としてとるとうまく行く分布) はベータ分布であるという話でしたが、多項分布のそれはディリクレ分布、正規分布のそれはある場合は正規分布、ある場合はガンマ分布、というようにいくつかのものが知られています。

第 8 章

決断理論と DNA 鑑定 2013 年 6 月

8.1 はじめに

DNA 鑑定において、ジェノタイプを調べて尤度比を求めたりするというのは、ある人が仮説を持っていて、その事前確率に差が小さいために決断できないときに、尤度比を提供することで、事後確率に変化させ、決断しやすくするためにやっていると言えます。この事前確率→情報提供→事後確率→事後確率に基づく決断という枠組みにこの章は基づいています。

8.2 決断

勉強会で使用したスライドはこちらです。

複数のオプションがあって、どちらかを選ばなくてはならないことはたくさんあります。服を選ぶ、レストランのメニューから料理を選ぶ…。ものすごく悩ましくて決められないこともあれば、即決即断できることもあります。また、同じ条件が与えられても、すぐに決める人もいれば、なかなか決めない人もいます。ここで「同じ条件」なのに個人差がある、ということが重要です。DNA 鑑定では、決断に資する情報を提供します。この情報は決断のための条件です。情報が同じ・条件が同じでも、個人によって選択行動は違います。では、情報以外に決断に影響している要因は何なのでしょうか？

このような問に答える、もしくは答えようとする理論が決断理論です。決断理論には、哲学、心理学、経済学、数学などが絡みます。DNA 鑑定とそこから出てくる尤度や確率

の情報は決断理論の中の数学 (の中の統計学) に含まれます。仮説間の尤度比が何倍だったら、決断できるか、というのは、心理学や哲学的な色彩を帯びてきます。特にその決断が「重大」なものであるとき、心理学・哲学は大きな意味を持つかもしれません。そもそも「決断が重大であるとは何か」というような哲学的な疑問も生じてきます。

8.3　決断・選択と生物の戦略

生物は太古の昔から、餌を探しにあっちに行こうかこっちに行こうかというような決断に迫られたり、繁殖戦略として有性生殖をしようか無性生殖にしようか、卵生にしようか胎生にしようか、というような決断・選択を繰り返してきました。そのような過程をへて、現在、生き延びている生物種は、「長い目で見たときに生き残る」戦略をとってきたものたちであろう、と考えられています。この「長い目でみて、サバイバルレースに勝ち残る戦略」として、情報科学が教える最適戦略は、「情報がないならないなりに、あるならあるなりに選択・決断すべし」「選択・決断にあたって、よさそうなものを選ぶのがよいが、少しは、よくないかもしれないものも選んでおくがよい」というものだと言うことです。このように選択・決断にばらつきを持たせるのがよいという戦略家の末裔である人間が、同じ条件を与えられたときに、決断にばらつきが出るのは、妥当かもしれません。

そんな視点から考えると、DNA 鑑定で提供するべきは、数学的に正確な情報であり、決断するべき人が使いやすい形で提供することが大事なのであって、その結果、各個人がどのような決断をするかには、『完璧な』介入はできないと考えることも適当かもしれません。

8.4　決断を支援するための３つのポイント

人が決断するとき、

- 事前分布
- 情報
- 事後分布

の３要素があります。法数学はこの３つのポイントのそれぞれで、役割を持つことができると思います。以下、ポイントごとに考えてみます。

8.5　事後分布の解釈法

決断するべき人の決断支援をするとして、仮説の尤度の違いとはどういうことかなどを理解可能な形式で説明することに尽きるように思います。判断については無色透明を目指すことが数学的には妥当かと思います。

8.6　事前分布の想定法

事前分布は「心の中」にあるという話が前の章で出てきました。心の中の事前分布に客観性を持たせるために情報を的確に活用する、という話もその際にしました。情報を的確に利用する、というのは言うのは簡単ですが、実行するのは非常に難しいです。ですから、難しいから、「心の中」の事前分布が個人個人で違っていてもよいことを認めた上で、その個人差を小さくする情報活用の仕方についての検討は重要であると思います。

具体的には、考慮に値する仮説の広さについて適切な推定をし、それを提示することは重要だと思われます。

また、ジェノタイプに基づく尤度・尤度比等が提供されたとき、事前分布が変われば、同じジェノタイプ情報でも事後分布は変わります。個人というのは、電卓ではないので、次のようにはできません。

「先日はこういう事前分布だった、そのときにジェノタイプ情報を使うと、これこれの事後分布になりますよ、と教えました。事前分布が変わったって？じゃあ、先日提供した尤度比を新しい事前分布に掛けてください、新しい事後分布が出ますから」

この計算ができるわけではないので、要請があるたびに、事前分布に違いが生じるたびに、ジェノタイプ情報を適用しなおして事後分布を解釈しやすい無色透明な説明で提示することが適当だろうと思います。

8.7　ジェノタイプ情報

DNA 鑑定に関する多くの数学的・統計学的テーマがこの部分に集中して当てはまります。特に気になることを挙げてみます。

取り扱いたい仮説が不特定多数を含む場合に、不特定多数の平均値や代表値を用いることが適切な場合があります。特に、尤度の期待値を計算するときなどは、それが適切であることも多いです。しかしながら、尤度の信頼区間をもとめたり、最大値・最小値を求め

たりする場合には、不特定多数が作っている分布自体を考慮しないとわからなくなります。極端な例としては、疑われている集団が赤の他人の場合とその中に血縁者が含まれている場合とがあると、集団の平均を用いることで失われる情報は後者でより大きくなる、というような例です。

そのほか、この範疇の課題として、勉強会にて教えていただいたものを以下に列挙します

- 試料が複数人の混合である場合
- 試料が希少で曖昧な実験結果しか得られない場合
- 実験結果のクオリティを考慮しなくてはならない場合
- 複数の試料がえられるも、整合性に欠ける場合
- 集団に関する情報が限局的である問題

ひとまず、本文書は、ここまでで一区切りとします！

第 II 部

番外編

第 9 章

DNA 多型情報を使って判断すること～仮説空間・割り付け問題・曖昧なデータについて～DNA 鑑定を題材に考える (DNA 多型学会講演)

9.1 はじめに

DNA 多型は、DNA 配列が同一種の中で一様でない部分・遺伝的多様性を持つ部分のことである。この遺伝的多様性は、同一種の個体間 (種内) にのみ認められるものではなく、異なる種の遺伝子配列も多様性を有している (種間の遺伝的多様性)。昨今、遺伝的多様性の実験技法として次世代シークエンサーが現れ、われわれが手にする遺伝的多様性の情報は大きく広がっている。その一つが、単一細胞シークエンス情報である。この技術により、個々の細胞の DNA 配列の違いを比較することができるようになっている。つまり、現在は、３段階の遺伝的多様性 (種間の遺伝的多様性、種内の遺伝的多様性、個体内の細胞別・遺伝的多様性) に関する情報が入手可能である。他方、次世代シークエンシング技術は、その高速かつ大容量の情報提供を特徴とするが、それとともに、配列情報の確率的・比率的情報提供という特徴も併せ持つ。この確率的・比率的情報というのは、ある意味では曖昧な情報とも言い換えられる。したがって、DNA 多型情報を用いたさまざまな判断作業 (遺伝因子マッピング、遺伝子診断、遺伝子配列に基づく分類、DNA 鑑定、など) において、曖昧さを有した情報を活用していくことが求められている。本講演では、次世代シークエンス技術に伴う DNA 多型情報の変化に触れつつ、曖昧な情報を用いた判

断について、「○か×か」という判断を求められることの多い DNA 鑑定を例に取って考察する。

9.2　本稿の全体構成

本稿では、DNA 多型と遺伝的多様性について述べた後、遺伝的多様性と DNA 鑑定との関係に触れ、実験と推定に関すること、微量試料・混合試料に関すること、仮説の組合せに関することについて順に検討する。検討にあたっては、DNA 鑑定の立場から検討したのちに、その医学・生物学研究における位置づけを確認する。

9.3　DNA 多型と遺伝的多様性

遺伝子多型・DNA 多型は大まかにサイズと配列変化パターンとで類別されている。大小様々な多型が存在し、この遺伝的多様性を捉えること自体が学問として成立するほど多彩である。遺伝子多型に見られる多様性の考え方を一般化することで、生物の多様性を構造的・体系的にとらえることが可能となるが、それによって生物の多様性は、種間の遺伝的多様性、種内の遺伝的多様性、集団内、個人内、発生・成長等の時間的な遺伝的多様性として階層的にとらえられる。このように様々な階層での遺伝的多様性について考えると、ある階層では「おおまかに均一であろう」とみなしていることを、別の階層では「不均一である」とみなしていることに気付く。種間の遺伝的多様性を考えているときには種内の多様性は無視する、というのが、この例である。このように、遺伝的多様性に関する実験結果というものは、ある見方をすると均一であるものも、別の見方をすれば不均一である、というような特徴がある。以降の項では、DNA 試料とその実験・実験結果には多様性・不均一性とが潜んでいること、また実験とは推定作業であるという観点から DNA 鑑定について考えていく。

9.4　DNA 鑑定と実験という推定作業

DNA 鑑定では個人の DNA 型の異同を問題にするものであるが、前項の考え方に照らすと、実際には、ある個人から異なる機会に採取した DNA 型が完全に一致することを前提としていることに気付く。そのためには、試料を完璧に取り扱い、かつ試料に含まれる DNA 分子の配列を実験という操作によって完璧に知ることが必要である。しかしながら実験とはあくまでも試料の状態から情報を取って推定する作業であり、「完璧」というこ

とはありえない。一般的な医学・生物学実験においては、この「推定ぶれ」が常に大きいことから、この点は逆に意識に上らせないことも多いが、DNA 鑑定では、実験結果としての DNA 型は「完璧」であると考えがちであるように思われる。しかしながら、実際には、条件のよい試料の場合に限ってもグレーゾーンの実験結果は必ず存在する。では、このグレーゾーンの情報はどのように扱われるべきであろうか。ひとつの簡易な方法は、グレーなものは切り捨てる、というやり方があるだろう。しかし、このやり方には注意する必要がある。切り捨てるということは、切り捨てなかった場合には「曖昧な結論」であったものが、切り捨てることによって、「確たる結論」に収束させている、ということであろう。この「結論の収束」が DNA 鑑定領域でとりわけて問題になるのは、DNA 鑑定では、「真実は A なのか B なのか」というように２つの仮説を対比してどちらかに決めることを目的としていることが多いことに由来するようだ。なぜなら、仮説 A を「確たる結論」とすることは、仮説 B に「確たる誤りと結論」づけることであって、「曖昧な結論」を「確たる結論」に切り替えることは、２つの仮説の間で大いなる不公平を生じてしまうからである。このように DNA 鑑定の立場からは、曖昧さの残る実験結果に対してセンシティブであるべし、というメッセージが聞こえてくる。このメッセージは、医学・生物学分野での実験結果の利用の動向とも呼応している。医学・生物学分野の実験もオミックス解析などの大規模データ科学分野を中心に不確かさを内包した実験結果が量産されるようになっている。その領域ではデータのクオリティコントロールと呼ばれるプロセスがあり、実際には一部のデータを足切りしている。この足切りをするかしないか、足切りをすると、どのような影響が出るのか、ということを問題にしなくてはならないのは DNA 鑑定の場合とまったく同じである。また、医学・生物学分野の実験からは、確率的・尤度的な利用を前提としたデータが算出されることも多くなっている。それらに対応するべく、確率的・尤度的データ解析技法の整備が進んでおり、今後は DNA 鑑定分野でも医学・生物学分野でもそのようなアプローチが盛んになることと思われる。

9.5　ヘテロ接合体のタイピングと微量試料

ホモ接合体とヘテロ接合体について考えてみる。ホモ接合体を均一であると考えれば、ヘテロ接合体は不均一であるとみなせる。またヘテロ接合体は不均一ではあるが、ある特別な不均一状態であって、1:1 という「ちょうど」な比率で不均一であるという特徴もある。今、ヘテロ接合体であることを実験的に確かめるという作業を考えてみる。ある個人が持つ DNA 分子のちょうど 50 この点について、DNA 鑑定においては、微量試料問題において特に問題となるが、ナノテクノロジー化の進んでいる一般の医学・生物学分野の

実験結果でも同様の注意が払われることが適切である。また DNA 鑑定においては試料が希少であるがために再実験ができないことも多く、再実験をしないことを前提に実験結果を利用して判断することが迫られることもある。この点は再現性を重んじる、いわゆる科学分野においては関係ない問題だろうか。実際には大規模データ科学分野では、実験のハイスループット化が進み、個々の実験は迅速かつ安価で実施されているが、その実験の一部について再実験をするとなるとハイスループット技術の恩恵は限定的となるのが一般的で、結果として再実験をするモチベーションは低下する。したがって、実験結果に曖昧さが含まれることを前提として、再実験せずに判断するという課題も、やはり DNA 鑑定と一般の医学・生物学に共通していることがわかる。

9.6　混合試料

DNA 鑑定では現場試料が複数の個人由来の混合試料であることがある。採取された DNA がお互いに異なる DNA 配列を持つ細胞集団に由来する状況である。ある採取試料に関して、何系統 (DNA 鑑定の場合には何人) の細胞が混じっているのかも未知である。また、系統ごとにどれくらいの割合で含まれているのかもわからない。さらには、それぞれの細胞系統間の DNA 配列の違いがどこにあるのかも不明である。以上はある特定の採取試料に関する不明事項であるが、ある現場から複数の試料を取ると、今度は試料間で系統の混合比に違いが生じてくる。DNA 鑑定ではこのような試料における DNA タイピング実験結果をどのように解釈するか、どのように用いるか、ということを議論している。医学・生物学分野でも混合試料がすでに研究対象になっている。たとえば癌ゲノム解析である。癌部由来組織からの細胞集団は癌細胞集団と非癌細胞集団の混合である。これを 2 系統の細胞の集団であるとみなして DNA 配列の違いを検討すれば、これは比較的単純な混合試料解析である。さらに詳しく解析することもある。癌細胞集団は変異・染色体異常を集積しながら増大していく。その過程に注目すると、癌細胞集団はそれ自体が DNA 配列的に不均一な集団であり、その不均一性は癌組織の成長に伴い空間的に分布を取る。このように癌細胞に集団遺伝学的なアプローチをする場合には、非常に複雑な混合試料問題となる。この複雑な混合試料を混合試料のまま解析するには、確率的なアプローチが必須となる。また、試料の混合性を実験的に回避することも可能であり、そのためにはセル・ソーティングやマイクロ・ダイセクションなどの技術による細胞の選り分けと、シングルセル・シークエンスなどの微量試料対応技術などが用いられる。このように微量化をする場合には、個々の試料からは混合性は排除されるが、どの細胞が拾われてくるかという過程での「採取ぶれ」の振れ幅が大きくなり、また、個々の細胞でのエラーも増大するから、

結果の総合的解釈には、やはり確率的な視点の導入が必須となる。このように、混合試料問題も DNA 鑑定といわゆる医学・生物学研究とには共通する解析・解釈上の課題がある。

9.7 仮説空間

最後に実験とは異なる点についても触れておきたい。DNA 鑑定では複数の個人と複数の試料との対応関係を一括して判断するという課題が発生することがある。複数の個人が巻き込まれた事故の際の個人の同定問題などである。この場合、複数個人の割り付け問題として多数の割り付けパターンの中から何かしら有意な情報を引き出すというデータマイニング問題を解いていることになる。これは生命科学における大規模データからのデータマイニングに共通する課題である。大規模なデータが有している真実は、ある意味では一つであるが、その全体として一つの真実を理解することは不可能であり、我々研究者はそこから意味のある多数のメッセージを一つ一つ取り出しているに過ぎない。そしてすべてのメッセージは相互に依存しており、どれか一つのメッセージを立てると別のメッセージが立ちにくくなることがある。このように DNA 鑑定における多仮説並立の問題は、自然科学分野にも普遍的に存在する「問題が絡み合って解きがたい状況」の模型のようにとらえることもできる。この点でも両分野の解析上の知見が相補いあって進んでいくことが大いに期待される。

9.8 おわりに

以上、統計遺伝学分野において、ゲノムやそのほかのオミックスデータの解析をする立場から、いくつかの DNA 鑑定に関するテーマについて検討し、それらの多くが、生命科学一般の解析・アプローチと深い関係にあることを指摘した。謝辞本発表にあたっては、DNA 鑑定とその解析上の問題について数多くの貴重なご助言を京都大学医学部法医学講座の玉木敬二先生はじめ講座の皆様から頂いたことに深謝いたします。

第 III 部

数学の程度が高めなこと・R の使い方など

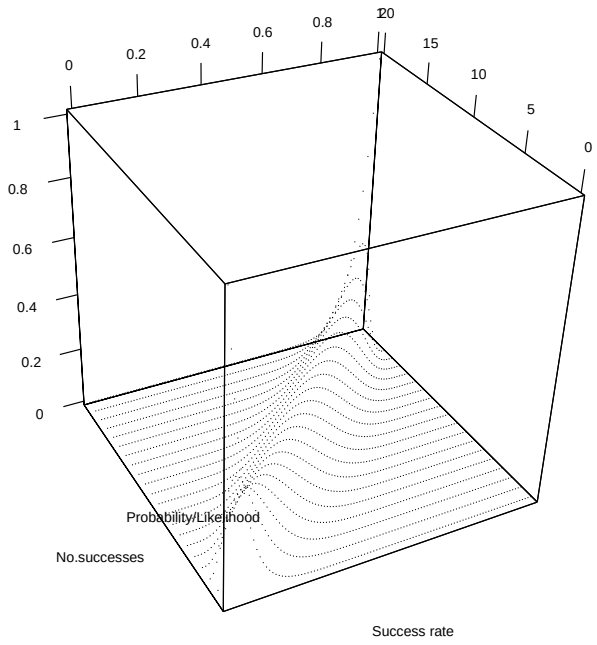
第 10 章

確率と尤度は同じものの見方が違うだけ

成功率が $0 \leq p \leq 1$ であるときに、20 回試行して、$k = 0, 1, 2, ..., 20$ 回成功する確率は

$$\binom{20}{k} p^k (1-p)^{20-k}$$

です。これを $p = 0, 0.01, 0.02, ..., 0.99, 1$ について計算してプロットしてみます。



No.successes は 0,1,2...,20 の 21 通りなので、21 本の山を持つ点線が表示されています。この点線の山が 20 回やって、k 回成功した場合の尤度です。一方、success rate の軸に着目すると、成功率を 0.01 刻みで 101 通り計算してプロットしてあります。ある特定の成功率に着目すると、k=0,1,2,...,20 についてそれぞれとびとびに 21 個の値を追いかけることができます。これが特定の成功率 p における k 回成功する生起確率です。このように確率と尤度は、同じ値を仮説 (成功率) の軸から眺めるか、結果 (成功回数) の軸から眺めるかの違いです

```
p <- seq(from=0,to=1,by=0.01)
N <- 20
k <- 0:N
pk <- expand.grid(p,k)
```

```
pr <- dbinom(pk[,2],N,pk[,1])
library(rgl)
plot3d(pk[,1],pk[,2],pr,
xlab="Success rate",ylab="No. successes",zlab="Probability/Likelihood")
rgl.postscript("problikelihood.eps")
```

第 11 章

最尤推定値 対数尤度関数の微分

最尤推定値は尤度関数・対数尤度関数が最大となるようなパラメタの値です。関数が最大となるとき、関数は極大値を取っていることを利用します。極大値では関数の一次導関数が 0 になっていることを利用します。

$$LL(\mathbf{p}|\mathbf{x}) = \ln \binom{+\sum_{i=1}^{6} x_i}{x_1, x_2, ..., x_6} \sum_{i=1}^{6} x_i \ln p_i$$

を微分します。$p_6 = 1 - \sum_{i=1}^{5} p_i$ なので $\frac{dp_6}{dp_i} = -1; i = 1, 2, ..., 5$ に気をつけて

$$\begin{aligned}\frac{\partial LL(\mathbf{p}|\mathbf{x})}{\partial dp_i} &= \ln \binom{\sum_{i=1}^{6} x_i}{x_1, x_2, ..., x_6} + \sum_{i=1}^{6} x_i \ln p_i \\ &= x_i \frac{1}{p_i} - x_6 \frac{1}{p_6} = 0\end{aligned}$$

従って

$$\frac{x_1}{p_1} = \frac{x_2}{p_2} = ... = \frac{x_6}{p_6}$$

であるから、

$$p_i = \frac{x_i}{\sum_{i=1}^{6} x_i}$$

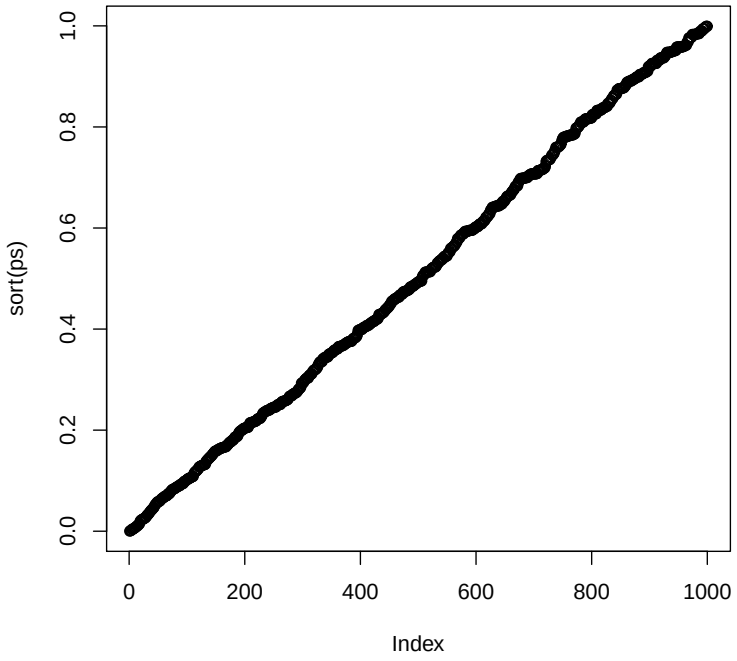
第 12 章

尤度比検定～サイコロの例～

理想的なサイコロを振って 6 つの目の出た数を観察する。その上で出た目に比例した確率 (最尤推定値) を対立仮説の生起確率ベクトルとして尤度比検定を行う。この試行を何度もくりかえすと尤度比検定の p 値は一様分布になることを示す。一様分布に従う値をソートしてプロットすると対角線上に並びます。

```
> # 理想的サイコロの目の出る確率
> p1 <- rep(1/6,6)
> # n.iter 回、試行試行する
> n.iter <- 1000
> # １試行あたりサイコロを N 回振る
> N <- 100
> # 各試行の p 値を格納するベクトル
> ps <- rep(0,n.iter)
> # 試行の繰り返し
> for(i in 1:n.iter){
+         # N 回のサイコロの目
+         s <- sample(1:6,N,replace=TRUE,prob=p1)
+         # カウントする
+         s.table <- tabulate(s)
+         # 対立仮説の最尤推定値
+         p2 <- s.table/N
+         # 帰無・対立仮説の対数尤度
+         LL1 <- lgamma(sum(s.table)+1)-sum(lgamma(s.table+1)) + sum(s.table*log(p1))
```

```
+         LL2 <- lgamma(sum(s.table)+1)-sum(lgamma(s.table+1)) + sum(s.table[which(s.table!=
+         # 対数尤度比の２倍
+         S <- 2 * (LL2-LL1)
+         # 自由度は５
+         df <- 5
+         ps[i] <- pchisq(S,df,lower.tail=FALSE)
+ }
> # p 値をソートしてプロット
> plot(sort(ps))
```



第 13 章

近親婚のない家系図でのジェノタイプ確率計算

13.1 地道な計算方法

- ある集団のジェノタイプ頻度情報が与えられている
- その集団において、ジェノタイプ情報つきの家系がどのくらいの頻度で存在しているかを計算する
- 独立した座位は座位ごとに計算して、積をとる
- 家系内には「集団から直接生まれた子」がいる
  - 家系内の「集団から直接生まれた子」とは、家系図で、親がない個人のこと。家系が「単名」の場合には、両親がいて、その両親が二人とも「集団から直接生まれた子」である場合に相当する
- 「集団から直接生まれた子」は、集団に存在しうるジェノタイプのすべてを取りうる
- そのような「集団から直接生まれた子」からスタートした家系の構成員のジェノタイプは、「集団から直接生まれた子」からアレルの伝達によって決まる
- あるジェノタイプ情報つきの家系が「集団の子」である確率は、「集団から直接生まれた子」から、アレルが伝達されて、観察しているようなジェノタイプを持つ構成員となる確率である
- 地道なアプローチ
- すべての個人は、すべてのジェノタイプを取る可能性があると考え、そのすべての場合について、「起きえるか」「起きえないか」、「起きるとしたら、それはどのくらいの確率で起きるか」を計算する

- 「起きるとしたら、どのくらいの確率」というのは、アレルの伝達が、50:50 であることを利用して計算する
- 50:50 でないのは「集団から直接生まれた子」のジェノタイプ頻度であって、それのみである
- そのようにして、すべての場合を確認した上で、「観察ジェノタイプ」に合致した場合を足し合わせればよい
  - 地道なので、わかりやすい
  - 大変だけれど、数え上げもいつかは終わる
  - 現実的な時間内には終わらない

13.2　確率計算。そのアプローチ、２つの比較

- 地道な方法は計算が終わらない
- 地道に計算しないで、同じ値を算出したい
- やり方１
  - 地道なやり方では、すべての個人がすべてのジェノタイプを取りうるというところからスタートした
  - 家系情報とジェノタイプが与えられると、ジェノタイプが確定している人は、そのジェノタイプしかとりえない
  - ジェノタイプが与えられていない個人については、家系図上の位置により、メンデルの法則を満足するジェノタイプしかとることができない
  - メンデルの法則を満足することを条件に、「取りうるジェノタイプ」を限定する
  - 限定されたジェノタイプの場合分けについて、網羅的に計算する
  - 計算するにあたって、核家族ごとに計算して、核家族をまたがった確率は、２つの（複数の）核家族に所属する個人のジェノタイプの場合分けにのみ留意して、確率を掛け合わせる
  - この方法がこちらの記事である
  - 核家族の連結をするというアルゴリズムなので、ループのある家系図に対応できない（ループがあるときは、そのループを含む範囲を一塊で取扱い、そのうえで、連結しながら計算することは可能なはず（ただし、面倒くさそうなので、実装していない）
- やり方２
  - 別のやり方で計算を省略する

- やり方１は２倍体ジェノタイプに関してメンデルの法則を適用して、場合の数を減らしたし、核家族の連結という考え方も、２倍体を基本とした取扱いである
- こちらのやり方２は、染色体に着目する
- 染色体に着目することで、要素数は２倍となる
- 要素数が増えれば、場合の数を考慮するべき対象も２倍となる
- その代わり、個々の要素の取りうる場合は減少するし、
- 個々の要素の間の関係が単純となるので、処理が単純になるメリットがある
- また、２倍体の場合分けでは、「AND、OR」といった処理が多いため、「数式」での確率計算に適さないのに対して、１倍体での場合分けは、その点もメリットがある
- さて、その具体的なやり方は次節で。

13.3　１倍体を単位として計算量を減らすアプローチ

- 概要
  - ハプロタイプグラフ
    - 家系図 (２倍体の伝達関係）から１倍体の伝達の様子（ハプロタイプグラフ）を作る
    - 家系図はすべての個人がつながっているが、ハプロタイプグラフはつながっていない
    - 母方と父方のグラフに分かれる
    - 「子はかすがい」
      - 「子が産まれると母方・父方のハプロタイプグラフが子を介して連結する」
  - 場合分け
    - 伝達パターンで場合分け
      - 親が子にアレルを伝えるときに、そのアレルが、その親の母方のものか父方のものかの場合わけ
      - こうすると、「枝分かれ」の無い、木グラフになっている
    - 母方・父方アレルに振り分ける場合分け
      - ２倍体のジェノタイプ（アレル２本分の情報）を母方アレル・父方アレルに振り分ける

  - メンデルチェック
    * アレルの伝達パターンとして確定したものが得られており
    * 父方・母方の伝達アレルも確定して考えているから
    * そのような伝達パターン・父母パターンが起こり得るかどうかの確認はとれる
      · 木グラフ上のすべてのアレルが同一であれば、「起こり得る」ことで、それ以外は「起こり得ない」こと
  - 木の観測確率
    * すべての木は、一つの「祖先染色体」を持つ
    *「祖先染色体」は、「集団からの直接の子」の染色体である
    *「祖先染色体」があるアレルを持つ確率は、（HWE を仮定すれば）集団のアレル頻度に等しいので、それが確率
    * この集団のアレル頻度と、「場合分けされた木の確率（これは、伝達分岐・父母パターンの割り付けが作る確率なので、$\frac{1}{2^k}$ の形をしている
    * したがって、木の観測確率は、アレル頻度と $\frac{1}{2^k}$ とでできている
  - ハプロタイプグラフ全体の観測確率
    * ハプロタイプグラフは、複数の連結グラフでできているから、個々の連結グラフを場合分けをにより、木グラフにしたうえで、確率を計算し、すべての連結グラフ（木）での確率の積をとり
    * それをすべての場合について足し合わせる
- ソースはこちら

13.4　１倍体で場合分け　ソース　ベータ版

```
# packages
# 依存パッケージ
library(kinship) # 家系図を描く
library(MCMCpack) # １倍体場合分けのみなら不要かも
library(gtools) # 場合分け　１倍体場合分けのみなら不要かも
library(sets) # 集合
library(paramlink) # 連鎖解析 核家族 # １倍体場合分けのみなら不要かも
library(rSymPy) # 数式演算
```

```
# ハプロタイプグラフを作る
## ハプロタイプの親子関係を作る(複数のグラフが混在)
MakeHaplotypeGraph2<-function(p){
        ns<-length(p[,1])
        ret<-matrix(0,ns*2,2)
        # 2*i-1,2*iはi番個人の母方と父方の染色体
        for(i in 1:ns){
                tmpmom<-p[i,2]
                tmpdad<-p[i,3]
                ret[2*i-1,1]<-2*tmpmom-1
                ret[2*i-1,2]<-2*tmpmom
                ret[2*i,1]<-2*tmpdad-1
                ret[2*i,2]<-2*tmpdad
        }
        ret[which(ret<0)]<-0
        ret
}

## ハプロタイプグラフを連結グラフ別に名前をつける

SeparateGraphs<-function(hG){
        ns<-length(hG[,1])
        assigned<-rep(0,ns)
        cnt<-1
        for(i in 1:length(hG[,1])){
                if(hG[i,1]!=0){
                        tmpself<-assigned[i]
                        tmpp1<-assigned[hG[i,1]]
                        tmpp2<-assigned[hG[i,2]]
                        M<-max(tmpself,tmpp1,tmpp2)
                        if(M==0){
```

```
                                M<-cnt
                                cnt<-cnt+1
                                assigned[c(i,hG[i,1],hG[i,2])]<-M
                            }else{
                                #M<-max(tmpself,tmpp1,tmpp2)
                                trioID<-c(i,hG[i,1],hG[i,2])
                                trio<-c(tmpself,tmpp1,tmpp2)
                                for(j in 1:length(trio)){
                                    if(trio[j]!=0){
                                        assigned[which(assigned
                                    }else{
                                        assigned[trioID[j]]<-M
                                    }
                                }
                            }
                            #print(assigned)

                        }
                    }
                    #assigned
                    G<-list()
                    cnt<-1
                    hG2<-cbind(1:ns,hG)
                    for(i in 1:max(assigned)){
                        if(length(which(assigned==i))!=0){
                            G[[cnt]]<-hG2[assigned==i,]
                            cnt<-cnt+1
                        }

                    }
                    G
```

```
}

# 伝達関係の場合分けを列挙
MakeDecsendPattern<-function(hG){
        Alt<-which(hG[,1]!=0)
        #Alt<-1:length(hG[,1])
        # その数
        nAlt<-length(Alt)
        # 選択の場合を列挙
        s<-expand.grid(rep(list(c(1,2)),nAlt))
        for(i in 1:(2^nAlt)){
                tmp<-matrix(0,nAlt,2)
                for(j in 1:nAlt){
                        tmp[j,1]<-Alt[j]
                        tmp[j,2]<-hG[Alt[j],s[i,j]]
                }
        }
        Ns<-length(hG[,1])/2
        Ls<-length(s[,1])
        s2<-matrix(0,Ls,2*Ns)
        cnt<-1
        for(i in 1:Ns){
                if(p[i,2]==0){
                        s2[,i*2-1]<-1
                        s2[,i*2]<-2
                }else{
                        s2[,i*2-1]<-s[,cnt]
                        cnt<-cnt+1
                        s2[,i*2]<-s[,cnt]
                        cnt<-cnt+1
                }
```

```
        }

        s2

}

# 二つのアレルを父方・母方に振り分ける場合分け

# ジェノタイプknownの個人の２アレルを
# paternal/maternalに割り付ける場合分けを列強する
# ホモの場合も２パターンを作る方が
# 計算は多くなるが、間違いが少なくて済みそうだ…
MakePatMatPattern<-function(g){
        GenotypePlus<-which(g[,1]!=0)

        heteros<-rep(2,length(GenotypePlus))
        heteros[which(g[GenotypePlus,1]==g[GenotypePlus,2])]<-1
        heterolist<-list()

        for(i in 1:length(heteros)){
                heterolist[[i]]<-1:heteros[i]
        }
        s3<-expand.grid(heterolist)
        s3
}
# 母方・父方の場合分けに応じてアレルを割り当て

AssignHaplotype<-function(g,s){
        Ns<-length(g[,1])
        hapg<-rep(0,Ns*2)
        GenotypePlus<-which(g[,1]!=0)
```

```
        for(j in 1:length(GenotypePlus)){
                #print(g[GenotypePlus[j],unlist(s[j]])
                hapg[GenotypePlus[j]*2-1]<-g[GenotypePlus[j],unlist(s[j])]
                hapg[GenotypePlus[j]*2]<-g[GenotypePlus[j],(unlist(s[j])-1.5)*(-1)+1.5]
        }
        hapg

}

# 伝達木(林)の作成
## ハプロタイプグラフ情報と伝達パターンを決めれば、複数の木(林)になる

SelectTrees<-function(sepG,v){
        ret<-list()
        for(i in 1:length(sepG)){
                tmp<-sepG[[i]][,1:2]
                for(j in 1:length(tmp[,1])){
                        tmp[j,2]<-sepG[[i]][j,v[tmp[j,1]]+1]
                }
                #print(tmp)
                tmp2<-rep(0,max(tmp))
                cnt<-1
                for(j in 1:length(tmp[,1])){
                        if(tmp[j,1]*tmp[j,2]!=0){
                                M<-max(tmp2[tmp[j,1]],tmp2[tmp[j,2]])
                                #print(M)
                                if(M==0){
                                        M<-cnt
                                        cnt<-cnt+1
                                        tmp2[tmp[j,1]]<-tmp2[tmp[j,2]]<-M
                                }else{
```

```
                                                tmp2[tmp[j,1]]<-tmp2[tmp[j,2]]<
                                        }
                                }
                        }
                        ret[[i]]<-tmp2[tmp[,1]]
                }
                ret
}

# 林のメンデルチェックと木の上のアレルを抽出
## 林の木の上のジェノタイプがすべて同じかどうかのチェックをし、木のアレ
# 木全体のメンデルチェック
MendelCheck<-function(sTrees,sepGraphs,hapg){
        ret<-TRUE
        ret2<-c()
        ret3<-c()
        for(i in 1:length(sTrees)){ # グラフごとのループ
        #print("---")
        #print(max(sTrees[[i]]))
                for(j in 1:max(sTrees[[i]])){ # グラフにある木をループ
                        # 木にある、"0"でないアレルはすべて同一である必
                        nodes<-sepGraphs[[i]][,1][which(sTrees[[i]]==j)
                        #print(hapg)
                        #print(nodes)
                        numanc<-length(which(sepGraphs[[i]][,2][which(s
                        tmp<-hapg[nodes][which(hapg[nodes]!="0")]
                        #print(tmp)
                        #print("length set")
                        #print(length(as.set(tmp)))
                        if(length(as.set(tmp))!=1){
                                if(length(as.set(tmp))>1){
```

```
                                        ret<-FALSE
                                    }
                                    #ret<-FALSE
                                    ret2<-c(ret2,"0")
                                    ret3<-c(ret3,1)

                                    #return(list(mendel=ret,alleles=ret2,nAncestor=ret3))
                                }else{
                                    #ret<-TRUE
                                    ret2<-c(ret2,hapg[nodes][which(hapg[nodes]!="0")][1])
                                    ret3<-c(ret3,numanc)
                                }
                            }
                        }
                        return(list(mendel=ret,alleles=ret2,nAncestor=ret3))
}

# マーカーごとに確率計算
## メンデルチェックを通った場合というのは、「あり得る」場合
## その確率は、「集団からの直接の子」が木のアレルを持つ確率と、場合分けの取り分による
## すべての場合を加算する
## 計算式も作る

ProbPerMarker<-function(p,g,As,Ps,hG,sepGraphs,s2,Expression=TRUE){
        cnt<-0
        s3<-MakePatMatPattern(g)
        if(Expression){
                Vars<-list()
                for(i in 1:length(As)){
                        Vars[[i]]<-Var(letters[i])
                        #print(Vars[[i]])
```

```
            }
            retExpression<-Var("0")
        }

        retProb<-0
        for(i in 1:length(s3[,1])){
            # Pat/Mat割り付けパターンでハプロタイプアレルを確定する
            hapg<-AssignHaplotype(g,s3[i,])
            for(j in 1:length(s2[,1])){
                # sepGraphsと伝達パターンから、木を作って取り出
                sTrees<-SelectTrees(sepGraphs,s2[j,])
                # 木全体のメンデルチェック
                MendelOK<-MendelCheck(sTrees,sepGraphs,hapg)

                # MendelOK$mendelがTRUEなら確率加算
                if(MendelOK$mendel){
                    #print(MendelOK)
                    cnt<-cnt+1
                    tmpexp<-Var(1/length(s2[,1]))

                    tmpProb<-rep(0,length(MendelOK$alleles)
                    for(k in 1:length(tmpProb)){
                        if(!MendelOK$alleles[k]=="0"){
                            aid<-which(As==MendelOK
                            #print(MendelOK$alleles
                            #print(aid)
                            if(Expression){
                                tmptmpexp<-Vars
                                if(MendelOK$nA
                                    for(l i
```

```
                                                                    }
                                                                }
                                                                tmpexp<-tmpexp*tmptmpexp
                                                            }

                                                            tmpProb[k]<-Ps[which(As==MendelOK$alleles[k

                                                        }else{
                                                            tmpProb[k]<-1
                                                        }

                                                }
                                                if(Expression){
                                                        retExpression<-retExpression+tmpexp
                                                }

                                                #print(tmpProb)
                                                #print(prod(tmpProb^MendelOK$nAncestor)/length(s2[,1]))
                                                retProb<-retProb+prod(tmpProb^MendelOK$nAncestor)/length(s2
                                        }
                                        if(!MendelOK$mendel){
                                                #print(sTrees)
                                                #print(MendelOK)
                                        }

                                }
                        }
                        #print(cnt)
                        if(!Expression) retExpression=Var("0")
                        list(prob=retProb,express=retExpression)
}
```

```
# マーカーごとに尤度比を計算する
## ２つの仮説の尤度を計算してその比を取る

CalcLikeRatioPerMarker<-function(p,g1,g2,searched,As,Ps,hG,sepGraphs,s2
        retProb1<-ProbPerMarker(p,g1,As,Ps,hG,sepGraphs,s2,Expression)
        #print("-----")
        retProb2<-ProbPerMarker(p,g2,As,Ps,hG,sepGraphs,s2,Expression)
        expres1<-retProb1$express
        expres2<-retProb2$express
        #print(expres1)
        #print(expres2)
        ProbGenPop<-1
        genExp<-Var(1)
        for(i in 1:length(searched)){
                tmpallele1<-g2[searched[i],1]
                tmpallele2<-g2[searched[i],2]
                P1<-Ps[which(As==tmpallele1)]
                P2<-Ps[which(As==tmpallele2)]
                #print(P1)
                #print(P2)
                tmp<-P1*P2
                tmpVar1<-Var(letters[which(As==tmpallele1)])
                tmpVar2<-Var(letters[which(As==tmpallele2)])
                tmpgenExp<-tmpVar1*tmpVar2
                if(P1!=P2){
                        tmp<-2*tmp
                        tmpgenExp<-2*tmpgenExp
                }
                ProbGenPop<-ProbGenPop*tmp
                genExp<-genExp*tmpgenExp
```

```
        }
        #print(genExp)
        #print(ProbGenPop)
        retProb1x<-retProb1$prob*ProbGenPop
        #print(retProb1$prob)
        #print(retProb2$prob)
        list(P1=retProb1x,P2=retProb2$prob,Pratio=retProb2$prob/retProb1x,expression=expres
}

# すべてのマーカーについて繰り返して結果を格納する
## pは家系情報
## gsはジェノタイプ情報
## searchedは被捜索者リスト
## poolidはsearchedがGpoolのどのジェノタイプに対応させるか
## Allelesはアレル名
## Probsはアレル頻度
## Express　は計算式を出すか出さないかのオプション

CalcLikeRatioAllMarker<-function(p,gs,searched,poolid,Gpool,Alleles,Probs,Express=TRUE){
        CLRAM<-list()
        ped<-MakePedigreeFromFamilyInfo(p)
        plot(ped)
        # ハプロタイプグラフを作る
        hG<-MakeHaplotypeGraph2(p)

        # ハプロタイプグラフを分ける
        sepGraphs<-SeparateGraphs(hG)

        # 伝達の場合分けを列挙する

        s2<-MakeDecsendPattern(hG)
```

```
        for(na in 1:length(Alleles)){
                As<-Alleles[[na]]
                Ps<-Probs[[na]]
                g<-gs[,,na]
                g2<-g
                #print(Gpool)
                g2[searched,]<-Gpool[poolid,,na]
                CLRAM[[na]]<-CalcLikeRatioPerMarker(p,g,g2,searched,As,

        }
        CLRAM
}

# 出力関数

PrintStringCLRPM<-function(CLR){
        LocusName<-c("D8S1179","D21S11","D7S820","CSF1PO","D3S1358","TH
        "D13S317","D16S539","D2S1338","D19S433","VWA","TPOX","D18S51","
        st<-paste("Locus","LR","Expression",sep="\t")
        st2<-paste("Locus","LR","Expression",sep="␣")
        print(st2)
        cprod<-1
        for(i in 1:length(CLR)){
                tmp<-paste(LocusName[i],CLR[[i]]$Pratio,sympy(CLR[[i]]$
                tmp2<-paste(LocusName[i],CLR[[i]]$Pratio,sympy(CLR[[i]]
                st<-paste(st,tmp,sep="\n")
                #print(st)
                #st<-st+tmp+"\n"
                print(tmp2)
                #st<-paste(st,tmp,collapse="\n")
```

```
                cprod<-cprod*CLR[[i]]$Pratio
        }
        st2<-paste("cumulative LR",cprod,sep="\t")
        st<-paste(st,st2,sep="\n")
        #print(st)
        print(st2)
        st
}
```

```
# 実行コマンド
## 式情報とかをため込むので、たくさんの家系で回すと重い
## 適宜、回し方は調整を要す

searched<-3
st<-""
for(fi in 1:6){
        p<-pedigrees[[fi]]

        gs<-genotypesFamily[[fi]]


        CLout<-CalcLikeRatioAllMarker(p,gs,searched,fi,Gpool[,,],Alleles,Probs,Express=TRUE
        st<-paste(st,PrintStringCLRPM(CLout),sep="\n")
}

st
```

第 14 章

ハンガリアン法による最適割り付け

行に行方不明者、列にご遺体をとり、行列 M の第 i 行 j 列には、第 i 行方不明者が第 j 遺体である確率が入っているとする。最適割り付けは、以下のようにすぐ出せます。ハンガリアン法と呼ばれるアルゴリズムです。こちらを参考に。

```
> k<-100
> library(clue)
> # 適当に行列を作ってやる
> M<-matrix(abs(rnorm(k^2)),k,k)
> # "max":最大になるように割り付け
> out<-as.vector(solve_LSAP(M,maximum=TRUE))
> # 結果を出力
> print(out)

 [1] 32  1 62 79 85 54 88 40 100 84 14 30 71 45 67 56 53
[18] 10 48 90 66 77 39 24 47 60 46 21 93 82 15 18 59  2
[35] 73 38 96 25 99 13 22 65 63 28 78 37 81 80  5  8 86
[52] 27 26 17 49 87 58  4 75 91 43 41  9 16 70 83 51 20
[69]  6 31 34 19 72 64 61 92 94 95 11 55 74 52 50 23 36
[86] 76 12 33 57 44 35 89 42  7 68 98  3 97 69 29

> # ここから、「最大値」を計算してみる
> VAL<-sum(diag(M[1:k,out]))
> # print(out) と同じ値
> print(VAL)
```

```
[1] 257.4066
```

謝辞

京都大学医学部法医学講座の玉木敬二先生、ならびに、大学院生の竹内ｘｘさん、真鍋翔さん、川合千尋さんには、多くのことを教えてもらい、非常に貴重なディスカッションの機会をいただきました。ここに感謝いたします。また、法数学勉強会にご参加いただきましたｘｘ先生をはじめとする科捜研・科警研の皆様には、退屈だったりわかりにくかったり、誤解したまま話したり、ということが多々ありましたこと、この場を借りてお詫びしますとともに、それにもかかわらず、さまざまなトピックについて一緒に勉強する時間を共有させていただきましたことを感謝いたします。