アレル頻度の推定

法数学勉強会

2018/04/21

京都大学医学研究科統計遺伝学分野

山田　亮

前回に引き続き

- 多アレル座位
  - Y染色体ハプロタイプのように多数のアレルが知られているときに標本から集団のアレル頻度を推定する話
  - 全部で X 本の標本。X= x1 + x2 + … + xk (xi >=1; k >=1) と観察されているものとする
  - 母集団には p1 + p2 + … + pk + … + pK = 1 (pi >=0)のように、全部でK >=k 種類のアレルが存在し、そのアレル頻度piがあるものとする

今回も
「これが正解」という
Take-home message
はなし

ディリクレ分布を使う

確率密度関数	$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i-1}$ ここで、$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}$ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$
期待値	$\mathrm{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$ $\mathrm{E}[\ln X_i] = \psi(\alpha_i) - \psi(\sum_k \alpha_k)$ (ここで、ψ(・)はディガンマ関数)
最頻値	$x_i = \frac{\alpha_i - 1}{\sum_{i=1}^{K}\alpha_i - K}, \quad \alpha_i > 1.$

ディリクレ確率密度
～尤度･事前分布･事後分布～

- (3,1)の観察、もしくは、(3,1,0)の観察
- ((3+1) + (1+1)) ! / ((3+1)!(1+1)!) p1^3 p2^1
- =6!/(4! 2!) p1^3 p2^1

- ((3+1)+(1+1)+(0+1))!/((3+1)!(1+1)!(0+1)!) p1^3 p2^1 p3^0
- =7! / (4! 2! 1!) p1^3 p2^1

- (3,1,0)の尤度は、(3,1)の7/6倍

確率密度関数

$$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

ここで、 $\mathrm{B}(\boldsymbol{\alpha})=\dfrac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}$

$\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_K)$

最頻ベクトルは同じ

最頻値	$x_i = \frac{\alpha_i - 1}{\sum_{i=1}^{K} \alpha_i - K}, \quad \alpha_i > 1.$

- (3,1)の観察、もしくは、(3,1,0)の観察
- 3/((3+1)+(1+1) – 2) = 3/4

- 3/((3+1)+(1+1)+(0+1)-3) = 3/4

- 「最頻」→「最尤推定」

期待値が違う

期待値	$\mathrm{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$ $\mathrm{E}[\ln X_i] = \psi(\alpha_i) - \psi(\sum_k \alpha_k)$ (ここで、ψ(・)はディガンマ関数)

- (3,1)の観察、もしくは、(3,1,0)の観察
- (3+1) / ((3+1)+(1+1)) = 4/6 = 2/3

- (3+1) / ((3+1)+(1+1)+(0+1)) = 4/7

- 第1アレルの頻度の (3,1,0)の場合の期待値は、(3,1)のそれより小さい　(中央寄りになる)

アレル種類数の想定を変えると

- (p1,p2) と (p1,p2,p3=0) とのように
- 「ゼロを立てるか立てないかが違うだけ」の違いがあるアレル頻度ベクトルの
  - ディリクレ分布の確率密度が変わる
  - 最頻ベクトル(最尤推定結果に対応する)は同じ
  - 期待ベクトル(期待値の推定結果に対応する)は変わる

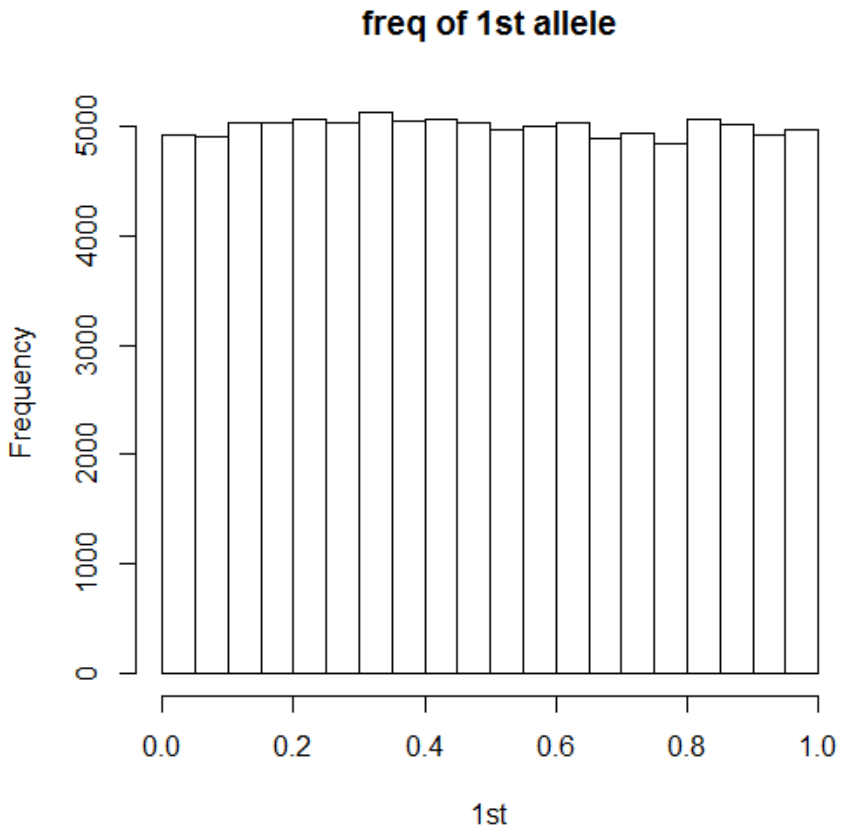
一様な事前分布はどんな分布か

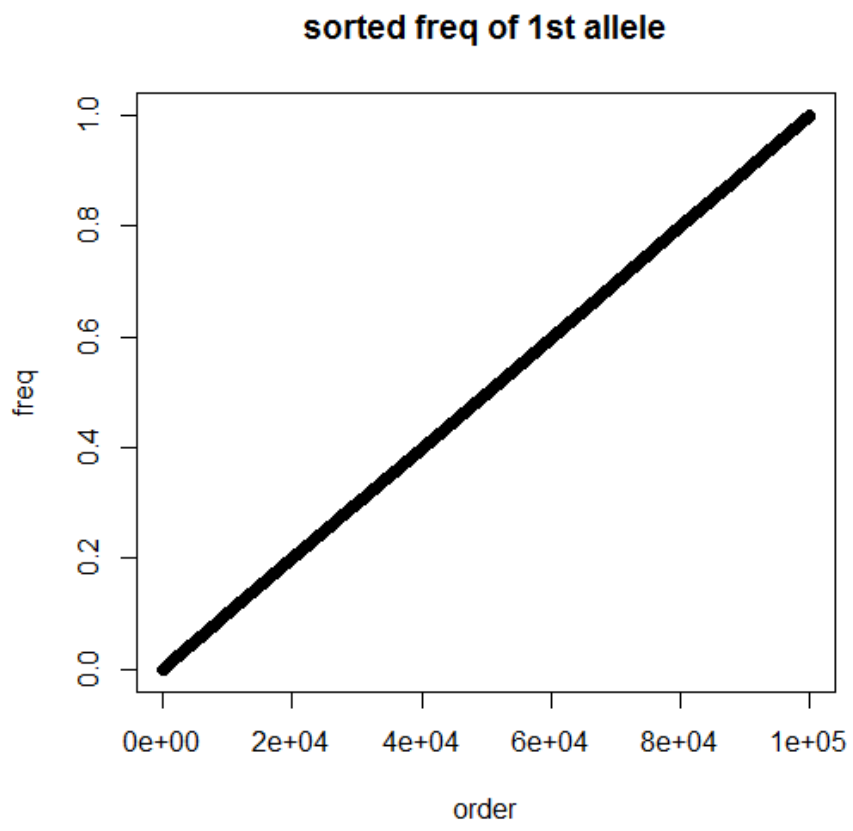
- ディリクレ分布のαは「集中度パラメタ」と呼ばれる
- α = (1,1,…)だと、確率密度関数はxの値によらない
- 言い換えると「平坦」な分布

確率密度関数	$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i-1}$ ここで、$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}$ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$
期待値	$\mathrm{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$ $\mathrm{E}[\ln X_i] = \psi(\alpha_i) - \psi(\sum_k \alpha_k)$ (ここで、ψ(・)はディガンマ関数)
最頻値	$x_i = \frac{\alpha_i - 1}{\sum_{i=1}^{K}\alpha_i - K}, \quad \alpha_i > 1.$

平坦な分布とはどういう分布か

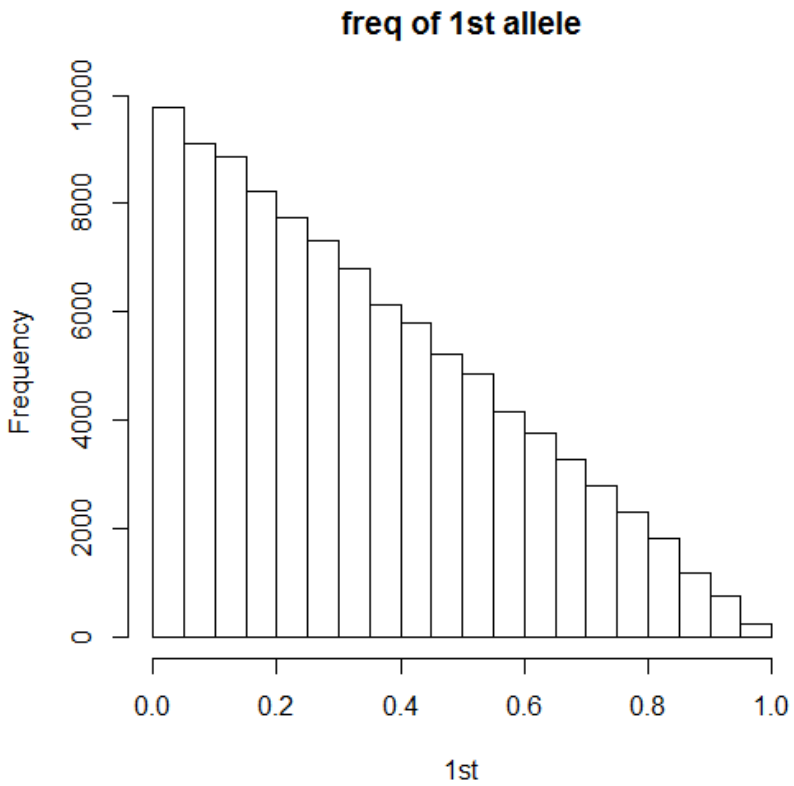
- ２種類の場合

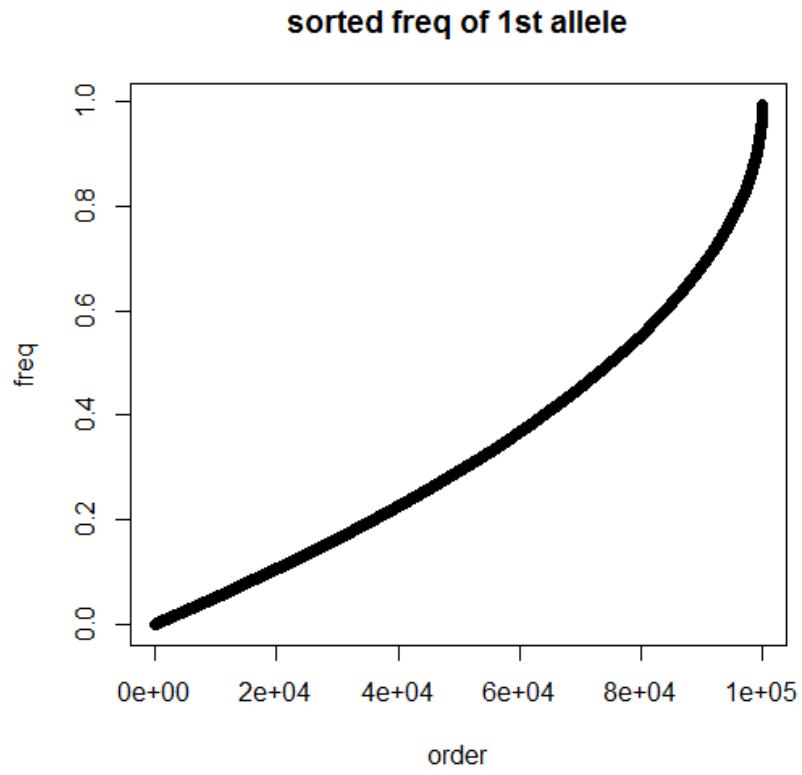




平坦な分布とはどういう分布か

- 3種類の場合

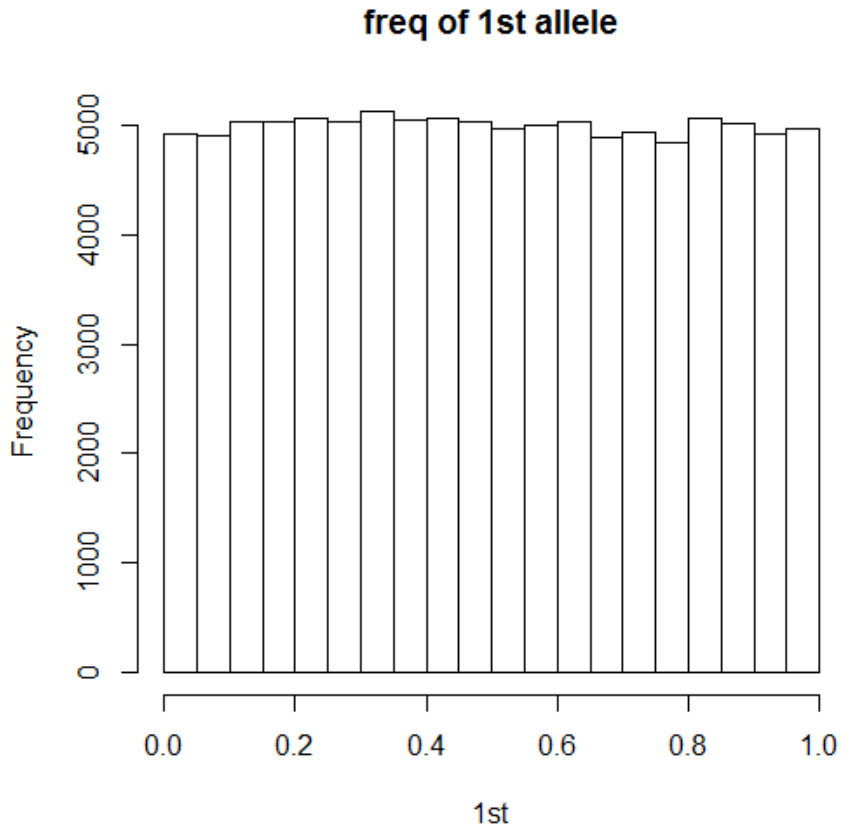




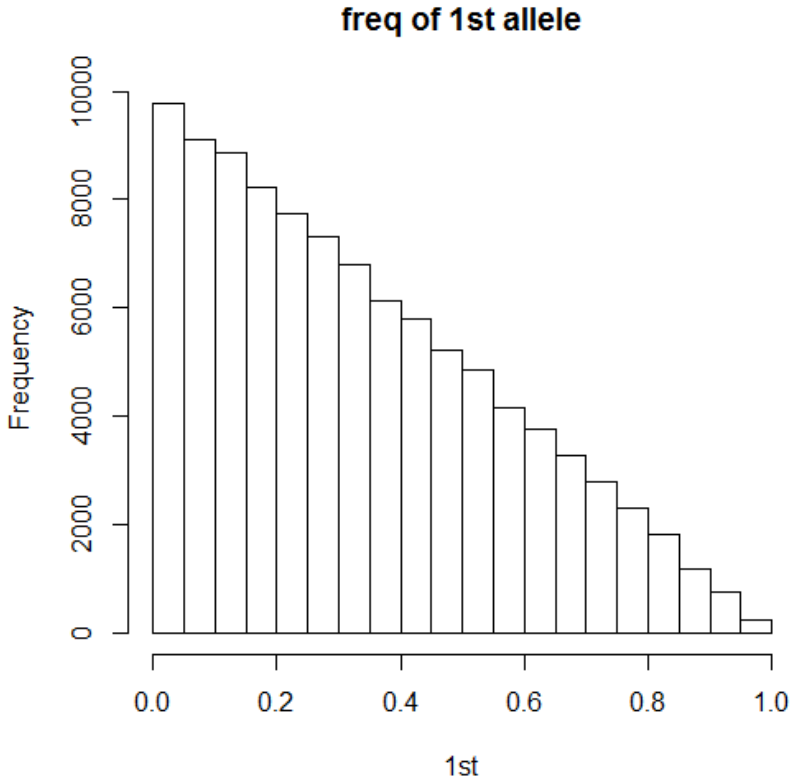
一様事前分布は
見方によって一様になったりならなかったり

- k = 3 の場合、特定のアレルの頻度は低い方が「好発」

k=2



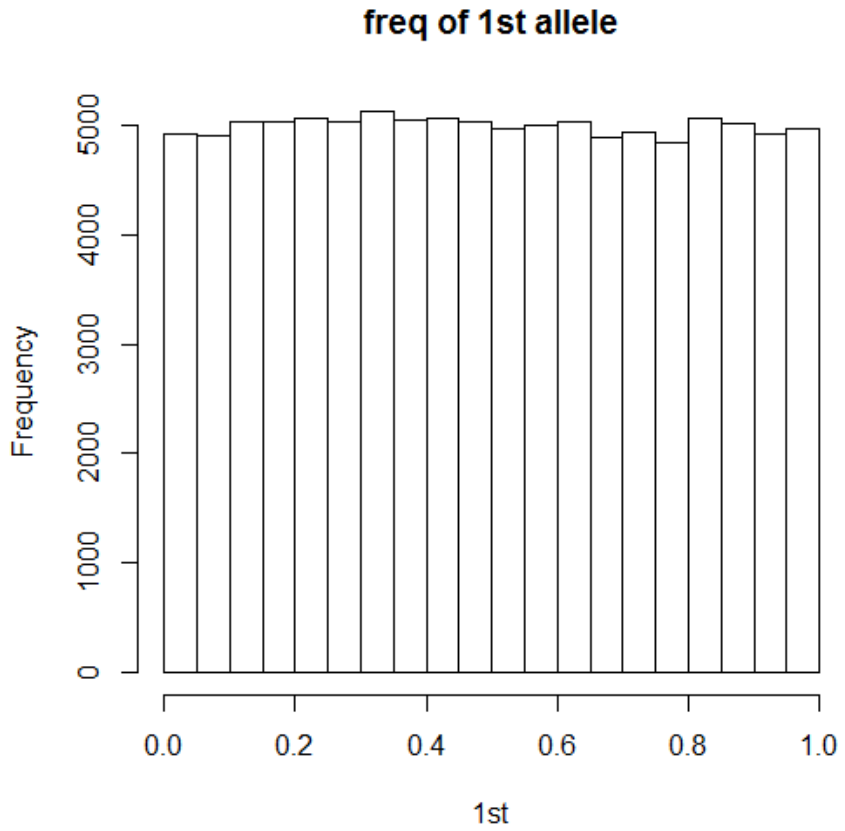
k=3



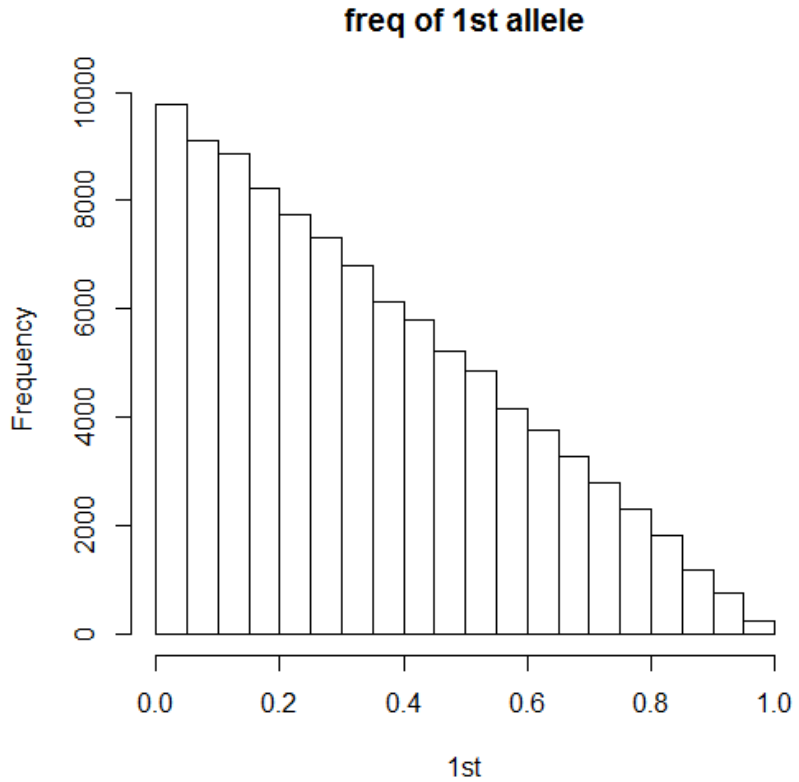
事前分布にディリクレ分布を使う場合

- k = 3 の場合、各アレルは低頻度のことが多いことを事前に想定している

k=2



k=3



もやもやするので、理解しやすい設定にする

離散条件で考える

離散で考える0あり/0なしの尤度

- 母集団の総数が決まっている場合
- 全部で5人、2種類しかない(1種類かもしれない、たかだか2種類)
- 母集団での人数が多い種類をT1、そうでない種類をT2とすれば
- (T1)=(5)または(T1,T2) = (4,1),(3,2)のいずれかに決まっている

離散で考える0あり/0なしの尤度

- 母集団の総数が決まっている場合
- 全部で5人、2種類しかない(1種類かもしれない、たかだか2種類)
- 母集団での人数が多い種類をT1、そうでない種類をT2とすれば
- (T1)=(5)または(T1,T2) = (4,1),(3,2)のいずれかに決まっている

- 2人を観察したところ、1種類が2人と観察されたという
- (T1)=5,(T1,T2)=(4,1),(3,2)のそれぞれの尤度はどのように計算すればよいだろうか？

離散で考える0あり/0なしの尤度

- 2人を観察したところ、1種類が2人と観察されたという
- (T1) = (5)の場合
  - 選び方の全通りは、1つ目の選び方は5通り、2つ目の選び方は4通り。
  - 観察に合致する選び方は、1つ目、2つ目、どちらも制限がないから、それぞれ5通り、4通り
  - したがって、尤度は (5 * 4)/ (5*4) = 20/20 = 1

離散で考える0あり/0なしの尤度

- 2人を観察したところ、1種類が2人と観察されたという
- (T1,T2) = (4,1)の場合
  - 選び方の全通りは、1つ目の選び方は5通り、2つ目の選び方は4通り。
  - 1つ目は、T1でないといけないので選び方は4通り。2つ目もT1でないといけないので、3通り
  - 尤度は、(4 * 3) / (5* 4) = 12/20

離散で考える0あり/0なしの尤度

- 2人を観察したところ、1種類が2人と観察されたという
- (T1,T2) = (3,2)の場合
  - 選び方の全通りは、1つ目の選び方は5通り、2つ目の選び方は4通り。
  - T1が2連続した場合とT2が2連続した場合とのそれぞれを考える必要がある
  - T1が2連続した場合は、3*2
  - T2が2連続した場合は、2*1
  - 尤度は、(3*2+2*1) / (5*4) = 8/20

離散で考える0あり/0なしの尤度

- 全5人、たかだか2種類
- 2人を観察したところ、1種類が2人と観察されたという

- 尤度を比較
- (T1) = (5) : 20/20 =1
- (T1,T2) = (4,1) : 12/20
- (T1,T2) = (3,2) : 8/20

離散で考える0あり/0なしの尤度

- 全5人、たかだか2種類
- 2人を観察したところ、1種類が2人と観察されたという

- 尤度を比較
- (T1) = (5) : 20/20 =1
- (T1,T2) = (4,1) : 12/20
- (T1,T2) = (3,2) : 8/20

ディリクレ分布の場合と比較する

離散で考える0あり/0なしの尤度

- ディリクレ分布の場合～尤度に比例した確率密度分布～
  - (3,1)の観察、もしくは、(3,1,0)の観察
  - ((3+1) + (1+1)) ! / ((3+1)!(1+1)!) p1^3 p2^1
  - =6!/(4! 2!) p1^3 p2^1

  - ((3+1)+(1+1)+(0+1))!/((3+1)!(1+1)!(0+1)!) p1^3 p2^1 p3^0
  - =7! / (4! 2! 1!) p1^3 p2^1

  - (3,1,0)の尤度は、(3,1)の7/6倍

離散で考える0あり/0なしの尤度

- ディリクレ分布の場合
  - 種類数を固定したときに、「全体を積分する」と1になる
  - 種類数が1かもしれず、2かもしれないとき
  - 種類数1でのディリクレ分布、種類数2でのディリクレ分布、どちらも積分すると1
- 離散の場合
  - (T1) = (5) : 20/20 =1　・・・合算して1・・・種類数1のディリクレ分布の離散版
  - (T1,T2) = (4,1) : 12/20
  - (T1,T2) = (3,2) : 8/20　・・・合算して1・・・種類数2のディリクレ分布の離散版

尤度は事前分布と併せて使う

- (T1) = (5) : 20/20 =1
- (T1,T2) = (4,1) : 12/20
- (T1,T2) = (3,2) : 8/20

- (T1)=(5),(T1,T2)=(4,1),(T1,T2)=(3,2)の事前確率をフラットに1/3,1/3,1/3 としたとすれば
- 3つの事後確率は、5:3:2 = 0.5:0.3:0.2

連続に戻して、考えてみる

- 『(T1)=(5),(T1,T2)=(4,1),(T1,T2)=(3,2)の事前確率をフラットに1/3,1/3,1/3 としたとすれば』
- ・・・離散での話
- 種類数1を想定したときに列挙できる場合と、種類数2を想定したときに列挙できる場合とに、フラットに事前確率を与えている
- ディリクレ分布で言うと:
  - 種類数3は正三角形
  - 種類数2は正三角形の辺
- 総人数をどんどん増やしていくと
  - 辺の長さはN+1
  - 三角形の面積はN(N+1)/2
  - 極限を取ると、面積の辺の場合数は面積に比べて無視できる小ささになる
  - ・・・これじゃあ、種類数が少ない場合の事後確率が0になってしまう・・・だめだ

離散でのフラットな事前分布を考え直す

離散で考えるフラットな事前分布1

- (T1)=(5),(T1,T2)=(4,1),(T1,T2)=(3,2)の事前確率をフラットに1/3,1/3,1/3 としたとすれば
- この仮定は離散ならあり、連続で困惑。

離散で考えるフラットな事前分布１

- 母集団の総数が決まっている場合
- 全部で５人、2種類しかない(1種類かもしれない、たかだか2種類)
- 情報なしのとき、何を事前に想定するか？
  - 「〇」か「×」かという2種類が存在すると知られていて
  - ５人の内訳が
    - (5,0),(4,1),(3,2),(2,3),(1,4),(0,5)の６通りのいずれかなのでは？
  - フラットな事前確率は、6通りのそれぞれについて1/6

離散で考えるフラットな事前分布１

- 母集団の総数が決まっている場合
- 全部で５人、2種類しかない(1種類かもしれない、たかだか2種類)
- 情報なしのとき、何を事前に想定するか？
  - 「〇」か「×」かという2種類が存在すると知られていて
  - ５人の内訳が
    - (5,0),(4,1),(3,2),(2,3),(1,4),(0,5)の６通りのいずれかなのでは？
  - フラットな事前確率は、6通りのそれぞれについて1/6

これの連続化は、そもそも1種類を考慮していない場合に相当する

離散で考えるフラットな事前分布１

- 母集団の総数が決まっている場合
- 全部で５人、2種類しかない(1種類かもしれない、たかだか2種類)
- 情報なしのとき、何を事前に想定するか？
  - 1種類のときはT1、2種類のときは現れた順にT1,T2と呼ぶことにする
  - ５人の内訳は
    - (T1,T2) = (5,0),(4,1),(3,2),(2,3),(1,4)の5通りになる
  - フラットな事前確率は、5通りのそれぞれについて1/5

離散で考えるフラットな事前分布１

- 母集団の総数が決まっている場合
- 全部で５人、2種類しかない(1種類かもしれない、たかだか2種類)
- 情報なしのとき、何を事前に想定するか？
  - 1種類のときはT1、2種類のときは現れた順にT1,T2と呼ぶことにする
  - ５人の内訳は
    - (T1,T2) = (5,0),(4,1),(3,2),(2,3),(1,4)の5通りになる
  - フラットな事前確率は、5通りのそれぞれについて1/5

これの連続化は、2種類の周辺である2点を合算しているだけで
相変わらず1種類の部分は、無限で消失する

離散で考えるフラットな事前分布1

- (T1)=(5),(T1,T2)=(4,1),(T1,T2)=(3,2)の事前確率をフラットに1/3,1/3,1/3 としたとすれば
- この仮定は離散ならあり、連続で困惑。

これの連続化は、コメントせずに来たけれど、「多い方をT1にする」的な処理により、ディリクレ分布の台(サポート)を折りたたんでいる。
前の例が(5,0)と(0,5)とだけ重ね合わせていたのに対し
(5,0): (0,5), (4,1) : (1,4), (3,2) : (2,3)をそれぞれ折り返している。

離散で考えるフラットな事前分布１

- 母集団の総数が決まっている場合
- 全部で５人、2種類しかない(1種類かもしれない、たかだか2種類)
- 情報なしのとき、何を事前に想定するか？
  - 種類は、1種類かもしれないし、2種類かもしれない
  - 1種類か2種類かにつき、フラットな事前確率としては、1種類であるという事前確率を1/2に、2種類であるという事前確率を1/2にすることもできる
  - このとき、1種類の内訳は、T1 = (5)。2種類である場合の内訳は(T1,T2)=(4,1)か(3,2)かである(母集団で多い方をT1、少ない方をT2とした場合)
  - 内訳についてのフラットな事前確率は (5) : 1/2、(4,1):1/4、(3,2):1/4

離散で考えるフラットな事前分布１

- 母集団の総数が決まっている場合
- 全部で５人、2種類しかない(1種類かもしれない、たかだか2種類)
- 情報なしのとき、何を事前に想定するか？
  - 種類は、1種類かもしれないし、2種類かもしれない
  - 1種類か2種類かにつき、フラットな事前確率としては、1種類であるという事前確率を1/2に、2種類であるという事前確率を1/2にすることもできる
  - このとき、1種類の内訳は、T1 = (5)。2種類である場合の内訳は(T1,T2)=(4,1)か(3,2)かである(母集団で多い方をT1、少ない方をT2とした場合)
  - 内訳についてのフラットな事前確率は (5) : 1/2、(4,1):1/4、(3,2):1/4

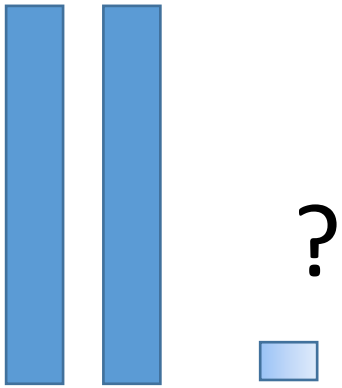
これは、あらかじめ、種類数ごとに事前確率を与えているので、これならば、少種類数の場合の事前確率が消えてしまわない

実例：離散で考える

- 医学科入学者100人の出身都道府県を考えます
- 47都道府県ごとの人数を x1,…,x47とします
- 100人に順番に出身都道府県を尋ねることにします
- t 番目の人の答えを知っているとき、t+1番目の人が答える都道府県が既出の都道府県のいずれかである確率をt番目までの情報を基に推定したい
- Y染色体のハプロタイプとほぼ同じ状況。ただし、47という上限値に相当するものがないのがY染色体の場合です

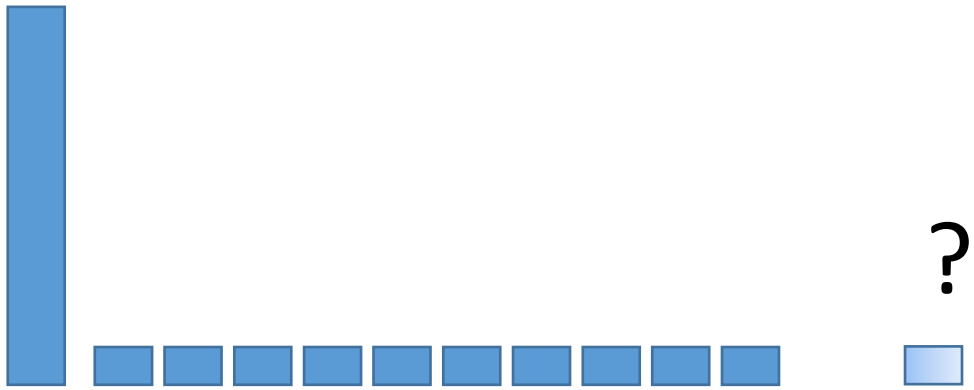
20人を調べたところ・・・

- (T1,T2) = (10,10)となったという
  - 21人目に未出現のタイプT3が現れるか、現れないか(T1,T2のいずれかになる)は、どちらがどれくらいありそうか？



20人を調べたところ・・・

- (T1,T2,T3,…,T11) = (10,1,1,…,1)となったという
  - 21人目に未出現のタイプT12が現れるか、現れないかは、どちらがどれくらいありそうか？



20人を調べたところ・・・

- (T1,T2) = (10,10)
  - 21人目に未出現のタイプT3が現れるか、現れないか(T1,T2のいずれかになる)は、どちらがどれくらいありそうか？
- (T1,T2,T3,…,T11) = (10,1,1,…,1)
  - 21人目に未出現のタイプT12が現れるか、現れないかは、どちらがどれくらいありそうか？

20人を調べたところ・・・

- (T1,T2) = (10,10)
  - 21人目に未出現のタイプT3が現れるか、現れないか(T1,T2のいずれかになる)は、どちらがどれくらいありそうか？
- (T1,T2,T3,…,T11) = (10,1,1,…,1)
  - 21人目に未出現のタイプT12が現れるか、現れないかは、どちらがどれくらいありそうか？

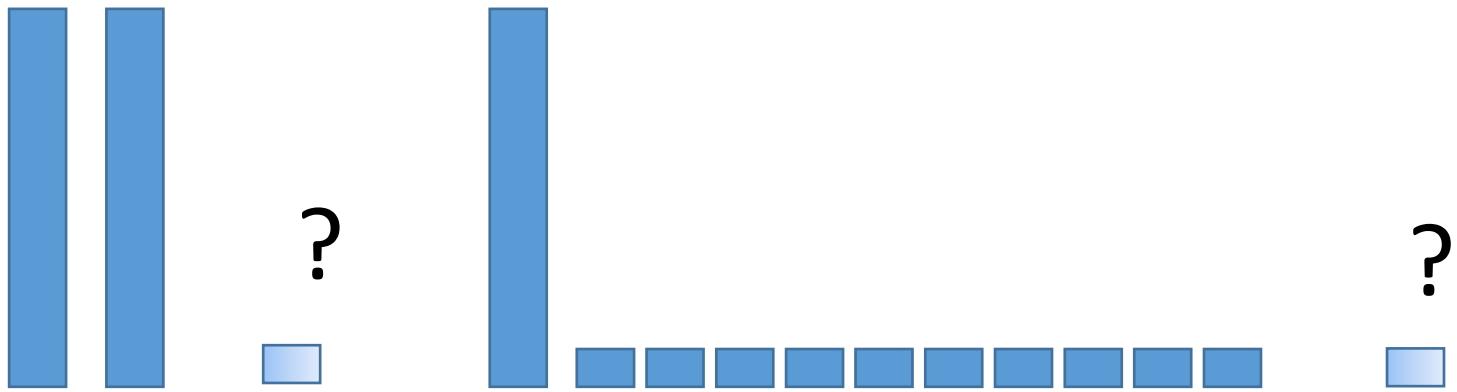
- どちらも、「未出現のタイプ」は20回連続して出現しないようなタイプであるという点で同じ
- 「既出現タイプ達」vs.「未出現タイプ」での２項対決をするなら、事後分布は同じはず～期待確率は同じはず

20人を調べたところ・・・

- (T1,T2) = (10,10)
  - 21人目に未出現のタイプT3が現れるか、現れないか(T1,T2のいずれかになる)は、どちらがどれくらいありそうか？
- (T1,T2,T3,…,T11) = (10,1,1,…,1)
  - 21人目に未出現のタイプT12が現れるか、現れないかは、どちらがどれくらいありそうか？

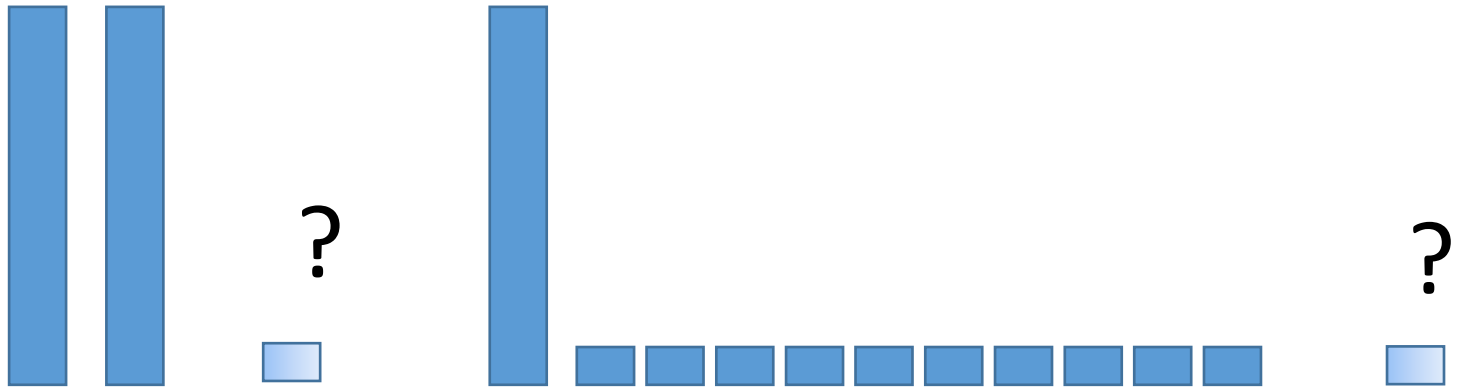
- どちらも、「未出現のタイプ」は20回連続して出現しないようなタイプであるという点で同じ
- 「既出現タイプ達」vs.「未出現タイプ」での２項対決をするなら、事後分布は同じはず～期待確率は同じはず
- 何が違うのか？

- (10,10)
- (10,1,1,1,1,1,1,1,1,1,1)



- (10,10)
- (10,1,1,1,1,1,1,1,1,1,1)

- 低頻度の県がたくさんあるなら、もっとあるのでは・・・という心理
- 個々の県(タイプ)の頻度は独立ではない・・・という心理



たとえば

- 頻度は指数関数型減衰に従うのでは、という心理～モデルを入れる、とか

今回も
「これが正解」という
Take-home message
はなし