アレル数が不明
～ディリクレ過程～

法数学勉強会

2018/07/21

京大(医)　統計遺伝学分野　山田亮

復習

- アレル数がわかっているとき
  - 2アレル(SNPとか)
    - 観測 : X本 vs. Y本
    - アレル頻度 (p, q = 1-p) の尤度 : ∝ p^X * q^(1-Y)
    - ベータ分布(k=2のディリクレ分布)
  - 3アレル
    - 観測 : X本, Y本, Z本
    - アレル頻度 (p, q, r; p + q + r = 1) の尤度: ∝ p^X * q^Y * r^Z)
    - k=2のディリクレ分布
  - kアレル
    - 観測: X1,….,Xk (Xi = 0 を含む)
    - アレル頻度 (p1, …, pk; p1 + … + pk = 1) の尤度: ∝ p1^X1 * … * pk ^Xk
    - k=kのディリクレ分布

復習2

- 観察 : X1 = 4, X2 = 3 のとき
- k = 2 ならk = 2 のディリクレ分布を使う
- k = 3 なら、X1 = 4, X2 = 3, X3 = 0 と考えて、k = 3 のディリクレ分布を使う
- k = 4 なら、X1 = 4, X2 = 3, X3 = 0 , X4 = 0 と考えて、k=4 のディリクレ分布を使う
- …
- k の想定によって、各アレルの頻度の期待値が変わる

母集団にアレル数が無限にある

- 無限にあるアレルのそれぞれが、0.000000001の割合だったら…
- 0.000000001 の∞倍 > 1 となりおかしなことになる
- 無限にあるアレルのいくつかは有限な頻度を持っているが、大多数は、頻度が”0”だけど、「存在はしている」、と考える
- これによって、実質的なアレル数は色々な数にできて、どんどん多くもできる

遺伝学では・・・

- 「生物採集をして、新種が見つかるかどうか」問題
  - “Probability of discovering new species”
  - 全部でK種いるのだろう（Kは不明）
  - N匹、採集したら、k (<= K) 種、観察された。
    - n1 + n2 + … + nk = N
  - 引き続き M 匹、採集したら、新種が s 種、発見される。その確率は？

Bayesian nonparametric estimation of the probability of discovering new species.
Antonio Lijoi et al. Biometrika(2007) 94,4, 769-786

何も仮定せずに「解ける」わけではないが…

- 種数（アレル数）無限の事前分布を何かしらモデル設定できれば
- そのモデルという仮定の下で
- 「生物採集をして、新種が見つかる確率」が
  - 正確に計算できる
    - 式を立てて解く
  - モンテカルロ・ベイズで推定できる
    - 「仮定」に基づいて事前分布を発生し、その尤度を計算できれば、モンテカルロ・ベイズでぐるぐる回せる
    - クラスタリング・分布推定などで使われる。「ノンパラメトリック・ベイズ」手法として括られていることもある

どんな仮定・どんなモデルか？

- 無限の種類数・アレル数を仮定しなければならない
- 大きく分けて、２つの考え方
  - サンプリングしたら、有限個数の多項分布が生じる
    - これなら有限個の標本観察の発生を無限種類数から発生させられる
    - 有限標本の生成に重きを置いている
  - 長さ1を分割する・無限分割することに関するもの
    - 母集団の種類比率は「足し合わせて１」を満足する必要がある
    - 母集団比率に重きを置いている

http://d.hatena.ne.jp/ryamada22/20180315

サンプリングしたら、有限個数の多項分布が生じる
中華料理店過程

定義 [編集]

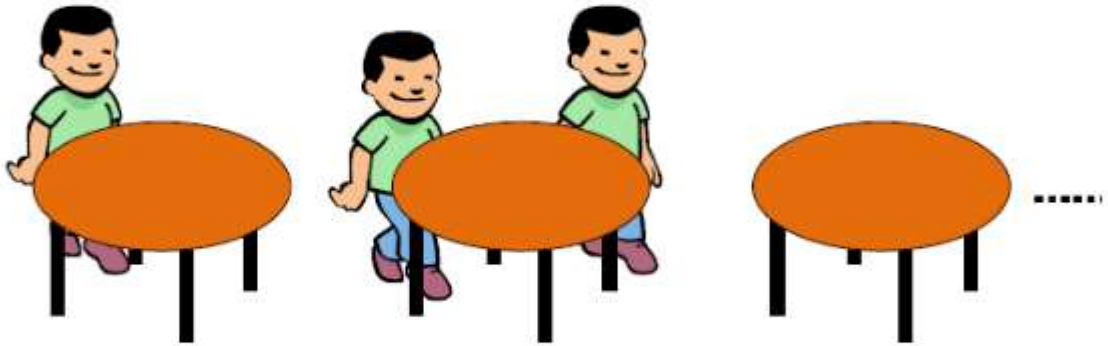
無限にたくさんの円卓が並べられた中華料理店を考える。各々の円卓もまた無限にたくさんの人が座ることが出来るものとする。1番目のお客が店に入ってくると、そのお客はまだ誰も座っていない円卓に確率1で座る。ある時刻$n+1$で現れる$n+1$番目の客は店内を見回し、より多くの人が座っている円卓に高確率で座ろうとする、あるいはまだ誰も座っていないテーブルに座ることもあるだろう。各々のテーブルが店にやってきた客の分割を与えるものだと考えたものが中華料理店過程の考え方である。

https://ja.wikipedia.org/wiki/中華料理店過程

中華料理店過程

- テーブルを選ぶ確率の総和が1になることを確認しよう
- 人数の多いテーブルが選ばれる確率が高いことを、式で確認しよう
- 未着席テーブルが選ばれる確率は、客の到着順とどのような関係にあるか考えよう

http://blog.datumbox.com/the-dirichlet-process-the-chinese-restaurant-process-and-other-representations/



$$P(z_i = c \mid c_{-i}) = \begin{matrix} 1 \\ \frac{1}{1+\alpha} \\ \frac{1}{2+\alpha} \\ \frac{1}{3+\alpha} \end{matrix} \quad \begin{matrix} 0 \\ \frac{\alpha}{1+\alpha} \\ \frac{1}{2+\alpha} \\ \frac{2}{3+\alpha} \end{matrix} \quad \begin{matrix} 0 \\ 0 \\ \frac{\alpha}{2+\alpha} \\ \frac{\alpha}{3+\alpha} \end{matrix}$$

http://d.hatena.ne.jp/ryamada22/20180320

中華料理店過程での「座り方」

一般化 [編集]

前述の中華料理店モデルは2つのパラメータαとθにより一般化できる。このときαとθはそれぞれ割引率と強度のパラメータと呼ばれる[1][2]。ある時刻n+1において新たに来店した客が|B|個のテーブルに人がいるのを確認して、まだ誰も座っていないテーブルに座る確率を、

$$\frac{\theta + |B|\alpha}{n + \theta}$$

とし、すでに|b|人が座っているテーブルに座る確率を

$$\frac{|b| - \alpha}{n + \theta}$$

https://ja.wikipedia.org/wiki/中華料理店過程

長さ1を分割する・無限分割する

- 無限に分割を繰り返す方法
- 有限回の分割で、無限分割をしたのと同じことにする方法

無限に分割を繰り返す方法

- ポアソン・ディリクレ過程
  - (0,1) に一様乱数を発生する （（一様）ポアソン・ディリクレ過程）
  - (0,1) に粗密を定めて、粗密に応じて乱数を発生する （非一様ポアソン・ディリクレ過程）
- 長さ1の線分の分割を繰り返す (Stick-breaking 過程）
  - 分割後の左側の線分はそのままにして、右側を分割し続ける （だんだん細かい線分が生じる）
  - ２分割の相対的位置はベータ乱数にする

有限回の分割で、無限分割をしたのと同じことにする方法

- Kingmanのpaintbox
- 何かしらの「無限分割法」を有限回行う
- １分割線分を「無限小」のタイプ数の束とみなす
- 「有限回処理」だけれど、「無限の種類数」が得られる

式で表せる場合

Ewens’s distribution

- n標本観察したら、k種類が観察された。biは第 i 番種類の観測標本数
- Θは、ある集団遺伝学的パラメタ

$$P(\Pi_n = \{b_1, b_2, \ldots, b_k\}) = \frac{\theta^k}{\theta(\theta+1)\ldots(\theta+n-1)} \prod_{j=1}^{k} (b_j - 1)!$$

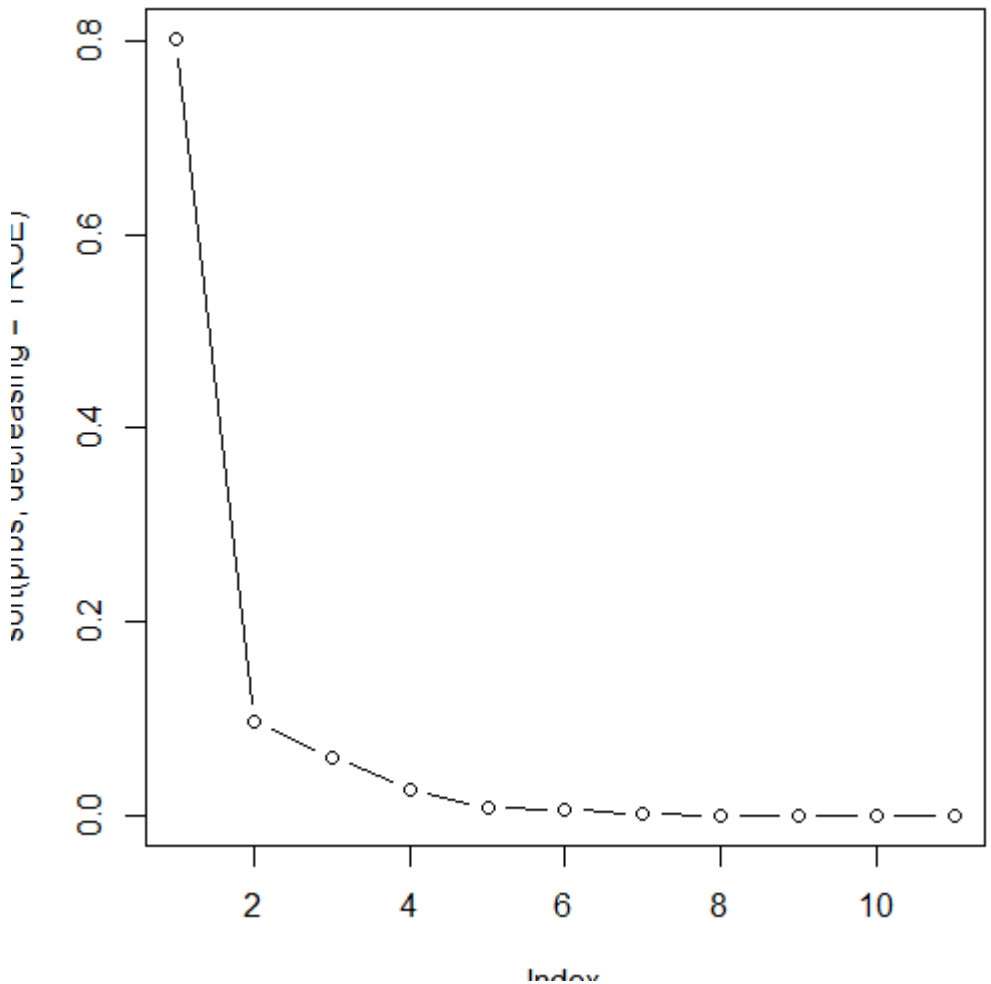
ちょっと背景があって、Ewens の式は別の表現が標準・・・

- Ewens's sampling formula
- n本の染色体を観察したとき、観察しうるアレルは1種類からn種類までである。また、アレル別に観察しうる染色体本数は1からnまでである
- 観測本数がj = 1,2,...,nであるアレル数をm_jとすると、$n=\sum_{j=1}^{n} j m_j$が成り立つ
- このとき、ある一定の条件を満足する集団においてn本の染色体の(n/2人：常染色体の場合)遺伝子座位を観察すると$(m_1,\ldots,m_n)$の生起確率は、あるパラメタθを用いて以下のようになることが知られている。ただし、θは有効集団サイズと変異率との積に比例した値である

$$P(m_1,\ldots,m_n|\theta)=\frac{n!}{\theta(\theta+1)\ldots(\theta+n-1)}\prod_{j=1}^{n}\frac{\theta^{m_j}}{j^{m_j}m_j!}$$

The Ubiquitous Ewens Sampling Formula.
Statist. Sci.
Volume 31, Number 1 (2016), 1-19.

https://projecteuclid.org/euclid.ss/1455115906



http://d.hatena.ne.jp/ryamada22/20180321

2. 中立モデル(遺伝)
    - Wright-Fisher model
        - 人口一定、ランダムメイティング、無限サイト、復元抽出
        - 遺伝子のモデルだが、確率モデルとしてはかなりがっちりとシンプルに徹しているモデル

- 遺伝学的使い道
  - 観察アレル種類数の確率分布も出せる(観察アレル数というのは、n=4のときに2本ずつ2種類、3本と1本との2種類というように、アレル種類数が同じならプールすること)
  - それがわかると、サイズNの母集団が有するアレル種類数と、サイズnの標本集団が有するアレル種類数についても分布がわかり、標本に観察されない未観測アレル種類数などについても予想がつけられる
  - Ewens's formulaからベタに計算することもできるし、第１種スターリング数を使った定義式$P(K=k)=|S_n^k|\frac{\theta^k}{\theta(\theta+1)...(\theta+n-1)}$でも計算できる

Rでやってみる…

パッケージ等が見つかりませんでした…

- すみません
- クラスタリング等のディリクレ過程仕様のパッケージは複数あるのですが
- 『新種の観測確率　～　Y染色体』に使えるものは見つかりませんでした

で、ディリクレ過程ってなんだったの？

- 「確率分布」を確率的に生成する過程

- 中華料理店過程は、テーブルを使って、多項分布を作り、極限として、母集団の無限種類分布を作った
- 分割法も、無限回分割で無限種類分布を作った

- 無限種類の分布を作れるので、それを事前分布として使ったベイズ推定ができる

参考資料等

- http://d.hatena.ne.jp/ryamada22/20180321　の前後
  - http://d.hatena.ne.jp/ryamada22/20180315
  - http://d.hatena.ne.jp/ryamada22/20180320
- http://didattica.unibocconi.it/mypage/dwload.php?nomefile=Bka_20072017021 0163749.pdf　新種発見確率のノンパラベイズ推定
- http://d.hatena.ne.jp/n_shuyo/20150626/dirichlet_process
- https://en.wikipedia.org/wiki/Dirichlet_process
- https://www.r-bloggers.com/dirichlet-process-infinite-mixture-models-and-clustering/
- http://statchiraura.blog.fc2.com/blog-entry-26.html
- http://www.stats.ox.ac.uk/~griff/pd.pdf
- http://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/pdf/1240-7.pdf