グレブナー基底で MCMC

@kyusque

グレブナー基底

- 計算機代数と実際のコンピューターの計算をつなぐ
- 僕もあまりよく分かってません
  - 数学にしては具体的でイメージしやすい（らしい）
- 取っ掛かりとしては、グレブナー基底大好きbotさんの本を読むといいと思います→
- 実は今回、グレブナー基底の計算はほぼ出ないです
  - (裏では計算しているみたいです)

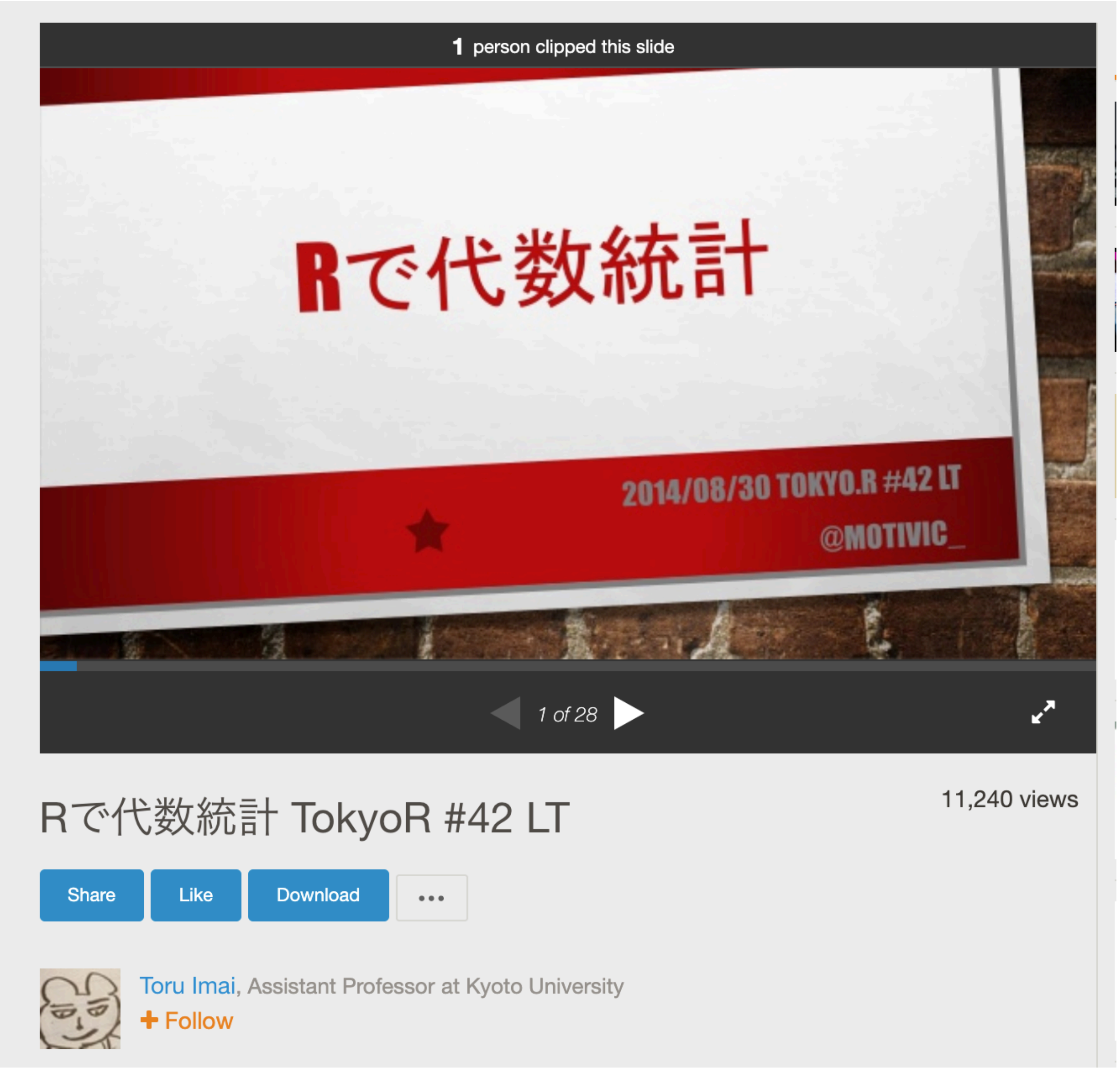
さらに学ぶために(統計スタート)

- グレブナー道場への入門書→
  - 数学科はCox, Little, O‘Shea の Ideals, Varieties, and Algorithms の第一章を英語で読むがお勧めだそうです
  - http://www.math.kobe-u.ac.jp/OpenXM/Math/dojo/dojo-mult/dojo-howto-read.html

- グレブナー道場へ入門するための入門書→

- 統計が分かる人はこちらの方が分かりやすい？→

実はRの先人が



- https://www.slideshare.net/motivic/tokyor42algstat

流れ

- ↓このPDFに沿ってやっていきます。
- マルコフ基底と実験計画（青木、竹村）
  - http://www.math.kobe-u.ac.jp/crest-c/cs-2012/2012-02-22-aoki-takemura.pdf

- あんまり自信がないので間違ってたらご指摘お願いします
- プログラムは以下に置いてます
  - https://github.com/kyusque/japanr2019

カイ二乗検定

胃ガンに対する喫煙のリスクを調べるため，胃ガン患者 20 人と健常者 100 人に対し，過去の喫煙習慣の有無を調査したところ，以下の結果を得た．喫煙習慣は胃ガンに関係あるといえるか?

	喫煙歴あり	喫煙歴なし	合計
胃ガン患者	14	6	20
健常者	56	44	100

(数値は架空)

- 行和、列和を固定して検定統計量のカイ二乗を計算
- カイ二乗分布に従うと仮定してp値を導出
  - 👆ここを代わりにMCMCで計算

http://www.math.kobe-u.ac.jp/crest-c/cs-2012/2012-02-22-aoki-takemura.pdf

カイ二乗検定

$\begin{matrix} + & - \\ - & + \end{matrix}$ の形の推移のみで十分（マルコフ基底をなす）

	喫煙歴あり	喫煙歴なし	合計
胃ガン患者	14	6	20
健常者	56	44	100

(数値は架空)

- グレブナー基底の計算で導出したマルコフ基底で行和、列和を固定して推移させていく

http://www.math.kobe-u.ac.jp/crest-c/cs-2012/2012-02-22-aoki-takemura.pdf

各種変換とマルコフ基底による推移

↓ここをMCMCで推移させる

14	6	20
56	44	100
70	50	120

↑行和、列和は固定

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} X = \begin{pmatrix} 20 \\ 100 \\ 70 \\ 50 \end{pmatrix}$$

A.mat

```
4 4
1 1 0 0
0 0 1 1
1 0 1 0
0 1 0 1
```

(4ti2)

```
$ markov A
```

👆裏でグレブナー基底を計算している(らしい)

A.mar

```
1 4
 1 -1 -1  1
```

$\begin{matrix} + & - \\ - & + \end{matrix}$ **の形の推移** →

計算の流れ

- 帰無仮説では、あるXが取り出される確率は超幾何分布に従う
- 観測されたXobsから得られる統計量χ2(Xobs)より、MCMCで得られるXから得られる統計量χ2(X)が大きい状況がどれくらいの頻度で起きるか(p値)
- 滅多になければ棄却

MCMCのアルゴリズム

アルゴリズムの例 (アルゴリズム 4.2.4)

入力: 観測値 $\boldsymbol{x}^o$, マルコフ基底 $\mathcal{B}$, 総ステップ数 N, 配置行列 A,

帰無分布 $f(\cdot)$, 検定統計量 $T(\cdot)$, ($\boldsymbol{t} = A\boldsymbol{x}^o$ とおく)

出力: p 値の推定値

変数: obs, count, sig, $\boldsymbol{x}$, $\boldsymbol{x}_{next}$

ステップ 1: $\mathtt{obs} = T(\boldsymbol{x}^o)$, $\boldsymbol{x} = \boldsymbol{x}^o$, $\mathtt{count} = 0$, $\mathtt{sig} = 0$

ステップ 2: $\mathbf{z} \in \mathcal{B}$ をランダムに, $\varepsilon \in \{-1, +1\}$ を等確率で選ぶ.

ステップ 3: $\boldsymbol{x} + \varepsilon\mathbf{z} \notin \mathcal{F}_{\boldsymbol{t}}$ であれば $\boldsymbol{x}_{next} = \boldsymbol{x}$ としてステップ 5 へ. $\boldsymbol{x} + \varepsilon\mathbf{z} \in \mathcal{F}_{\boldsymbol{t}}$ であれば u を 0 と 1 の間の一様乱数とする.

ステップ 4: $u \leq \dfrac{f(\boldsymbol{x} + \varepsilon\mathbf{z})}{f(\boldsymbol{x})}$ であれば $\boldsymbol{x}_{next} = \boldsymbol{x} + \varepsilon\mathbf{z}$ としてステップ 5 へ. $u > \dfrac{f(\boldsymbol{x} + \varepsilon\mathbf{z})}{f(\boldsymbol{x})}$ であれば $\boldsymbol{x}_{next} = \boldsymbol{x}$ としてステップ 5 へ.

ステップ 5: $T(\boldsymbol{x}_{next}) \geq \mathtt{obs}$ であれば $\mathtt{sig} = \mathtt{sig} + 1$

ステップ 6: $\boldsymbol{x} = \boldsymbol{x}_{next}$, $\mathtt{count} = \mathtt{count} + 1$

ステップ 7: $\mathtt{count} < N$ であればステップ 2 へ

ステップ 8: p 値の推定値は $\mathtt{sig}/N$

http://www.math.kobe-u.ac.jp/crest-c/cs-2012/2012-02-22-aoki-takemura.pdf

```{r}
mcmc(data_, move, 100000, chi_square, hypergeometric_dist)
```

```
[1]  9 11 61 39
[1] 14  6 56 44
[1] 11  9 59 41
[1] 12  8 58 42
[1] 13  7 57 43
[1] 13  7 57 43
[1] 11  9 59 41
[1] 10 10 60 40
[1]  9 11 61 39
[1]  8 12 62 38
[1] 12  8 58 42
[1]  9 11 61 39
[1] 15  5 55 45
[1] 13  7 57 43
[1] 13  7 57 43
[1] 14  6 56 44
[1]  7 13 63 37
[1] 12  8 58 42
[1] 16  4 54 46
[1] 12  8 58 42
[1] 14  6 56 44
[1]  9 11 61 39
[1] 14  6 56 44
[1] 10 10 60 40
[1]  9 11 61 39
[1] 10 10 60 40
[1] 13  7 57 43
[1] 12  8 58 42
[1] 10 10 60 40
[1] 11  9 59 41
[1] 10 10 60 40
[1]  9 11 61 39
[1] 14  6 56 44
```

```
[1] 10 10 60 40
[1] 11  9 59 41
[1] 15  5 55 45
[1] 12  8 58 42
[1] 12  8 58 42
[1] 10 10 60 40
[1] 13  7 57 43
[1] 15  5 55 45
[1]  9 11 61 39
[1] 11  9 59 41
[1] 16  4 54 46
[1] 12  8 58 42
[1]  8 12 62 38
[1]  9 11 61 39
[1] 13  7 57 43
[1] 14  6 56 44
[1] 13  7 57 43
[1] 16  4 54 46
[1] 15  5 55 45
[1] 11  9 59 41
[1] 12  8 58 42
[1] 13  7 57 43
[1]  8 12 62 38
[1] 15  5 55 45
[1] 12  8 58 42
[1]  8 12 62 38
[1]  8 12 62 38
[1] 14  6 56 44
[1] 14  6 56 44
[1] 11  9 59 41
[1] 12  8 58 42
[1] 13  7 57 43
[1] 13  7 57 43
[1] 0.32363
```

```
[1] 10 10 60 40
[1] 11  9 59 41
[1] 15  5 55 45
[1] 12  8 58 42
[1] 12  8 58 42
[1] 10 10 60 40
[1] 13  7 57 43
[1] 15  5 55 45
[1]  9 11 61 39
[1] 11  9 59 41
[1] 16  4 54 46
[1] 12  8 58 42
[1]  8 12 62 38
[1]  9 11 61 39
[1] 13  7 57 43
[1] 14  6 56 44
[1] 13  7 57 43
[1] 16  4 54 46
[1] 15  5 55 45
[1] 11  9 59 41
[1] 12  8 58 42
[1] 13  7 57 43
[1]  8 12 62 38
[1] 15  5 55 45
[1] 12  8 58 42
[1]  8 12 62 38
[1]  8 12 62 38
[1] 14  6 56 44
[1] 14  6 56 44
[1] 11  9 59 41
[1] 12  8 58 42
[1] 13  7 57 43
[1] 13  7 57 43
[1] 0.32363
```

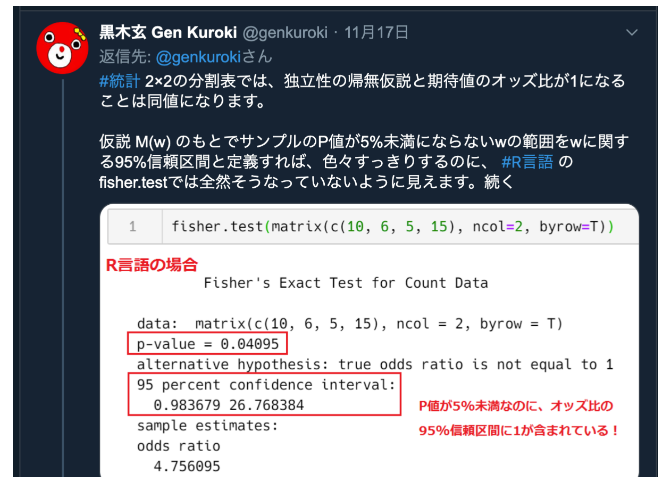
```{r}
chisq.test(data)
```

```
	Pearson's Chi-squared test with Yates' continuity cor

data:  data
X-squared = 0.82971, df = 1, p-value = 0.3624
```

まとめ

- グレブナー基底(マルコフ基底)から計算した遷移でカイ二乗検定のp値をMCMCで計算できた。
  - カイ二乗検定の妥当性についてはJulia言語界隈などで激論されている？みたいなので、今後も勉強していきたい



- 今回は遷移が一つだったが、表がより複雑になると遷移の数が増える
  - 3x3x3分割表などでも遷移を導出できる
- 実験計画法にも使えるらしい