

ベイズ推定の概要

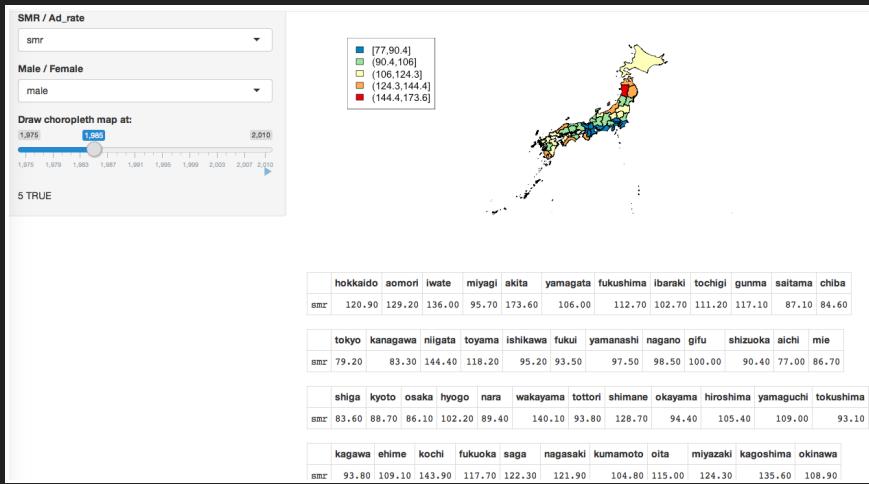
超入門

@広島ベイズ塾

Created by Yoshitake TAKEBAYASHI / @psycle44

自殺の時空間疫学

<http://ikiru.ncnp.go.jp/ikiru-hp/genjo/toukei/index.html>



自己紹介

名前: 竹林由武

所属: 統計数理研究所

専門: 疫学、臨床心理学

研究テーマ:

自殺の時空間疫学

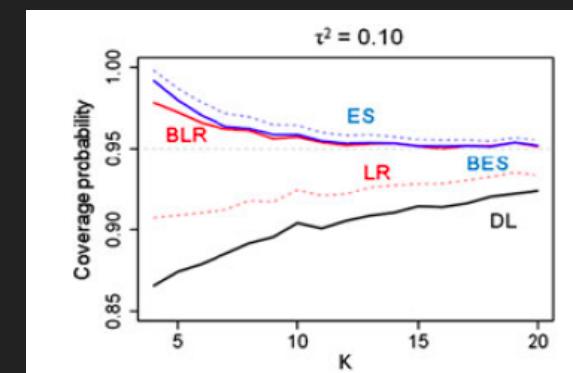
小研究数メタ分析

ウェルビーイング向上による精神疾患予防

猫の気持ち

少研究数のメタ分析

Noma, H. Statist. Med. 2011, 30 3304–3312



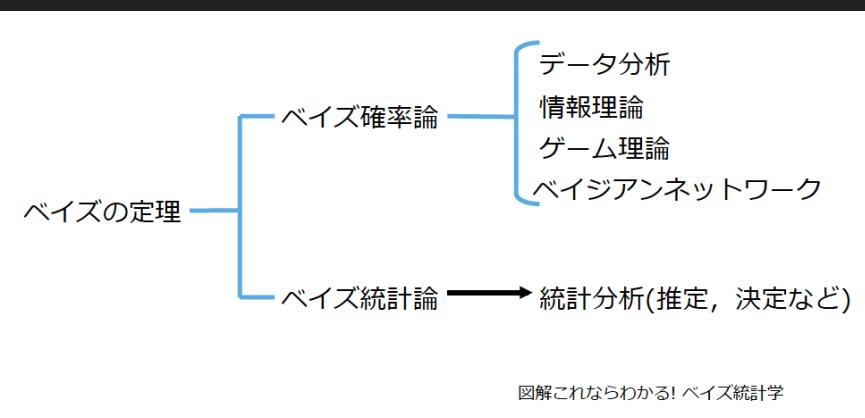
K=10以下の信頼区間を向上させる

TOPICS

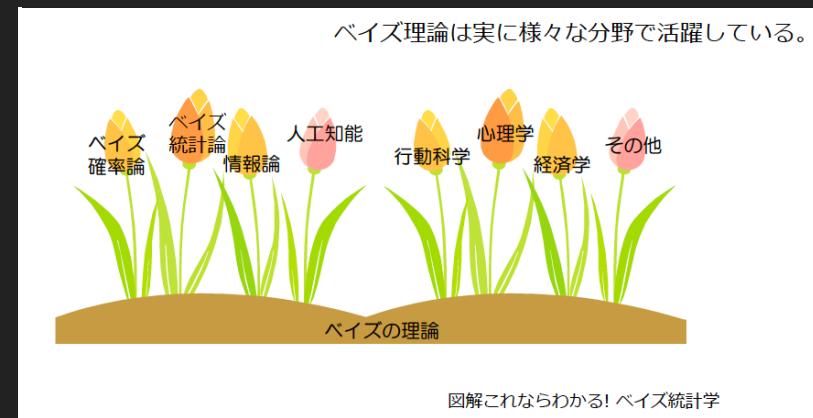
- はじめに
- 推測統計の基本
- 最尤推定とベイズ推定
- MCMCによるベイズ推定



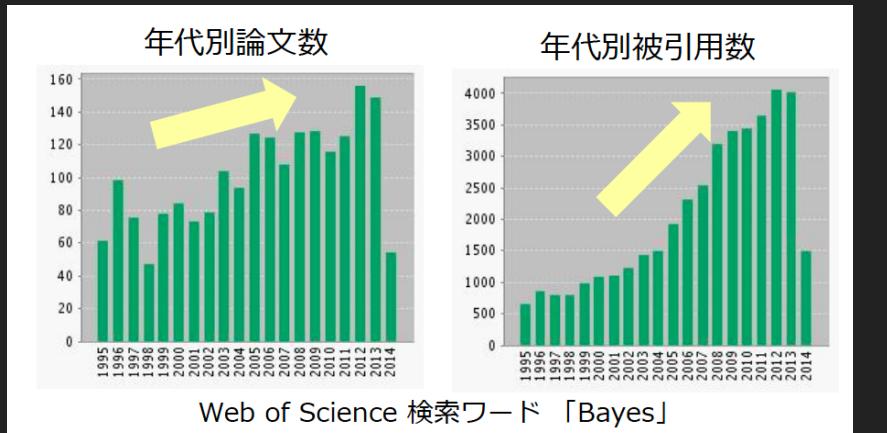
21世紀におけるベイズの発展



21世紀におけるベイズの発展



21世紀におけるベイズの発展



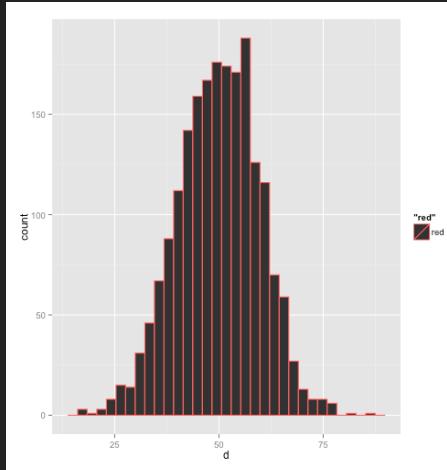
なぜ21世紀？？

ベイズ推定では
膨大な反復計算が必要
PCの処理性能の向上でベイズ推定
が実用的に!!

ベイズ推定のメリット
複雑なモデルを推定可能
少サンプル数でも推定可能
従来型の推定法も表現可能

要するに
めっちゃ便利!!
使えると解析の幅が広がる

ある地域での模試の得点の分布



関心:日本全体での分布が知りたい

ベイズ推定を始める前に....

推測統計学のおさらい

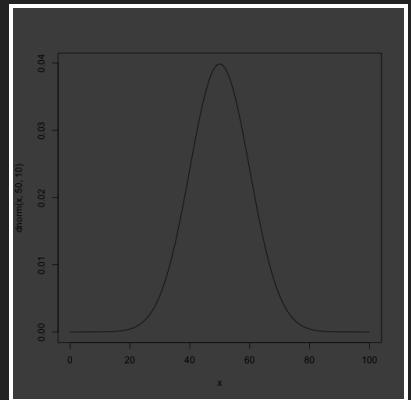
日本全体での模試得点の分布

統計モデル:正規分布

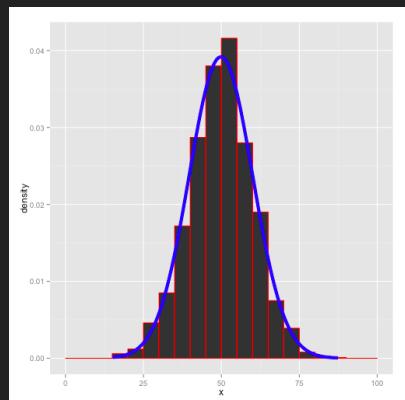
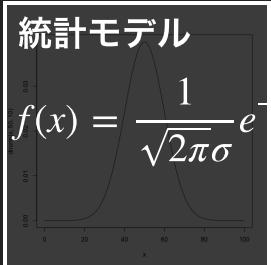
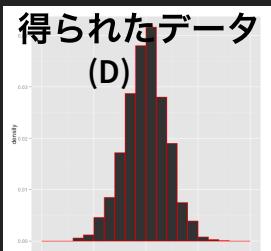
$$N(\theta, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

平均(θ)と分散(σ^2)
で分布の型が決まる



統計的推測



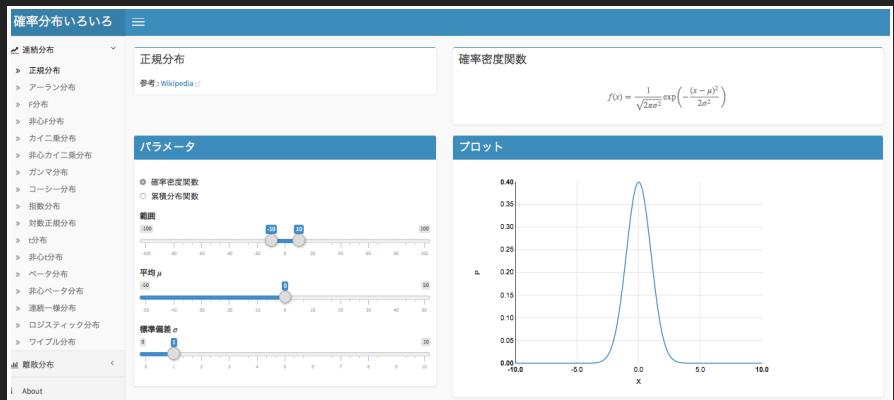
データに最もフィットする
統計モデルのパラメータを探す

データの分布型は多様

データにフィットする分布設定が
統計モデリングの要

様々な分布を知るために

<https://ksmzn.shinyapps.io/statdist/>



最尤推定とベイズ推定

どちらもパラメータ推定法だけど...

最尤推定法の発想

$P(D|\theta)$ が最大になる θ を推定する

$P(D|\theta)$ =ある**母数(θ)**の下で**データ(D)**が得られる確率

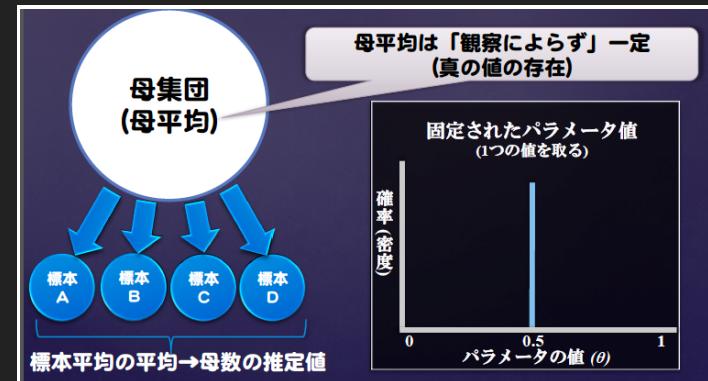
ベイズ推定法の発想

$P(\theta|D)$ を推定する

$P(\theta|D)$ =ある**データ(D)**の下で**母数(θ)**が得られる確率

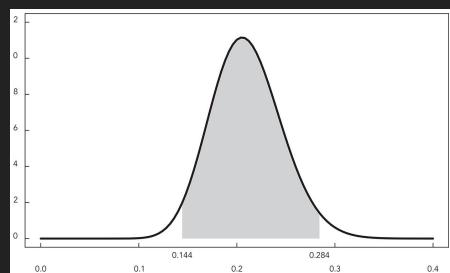
最尤推定法の発想

$P(D|\theta)$ が最大になる θ を推定する



最尤推定法の発想

信頼区間の意味

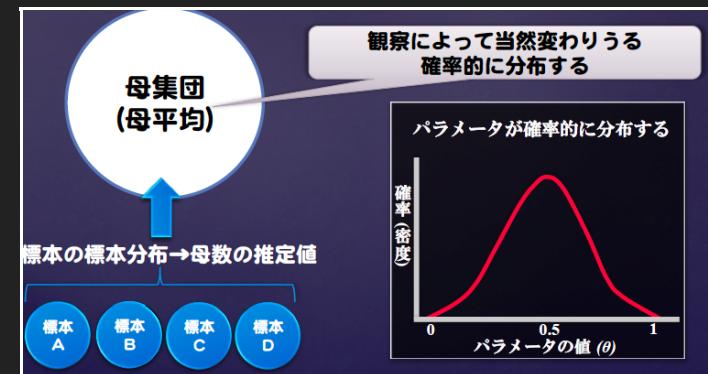


何度も観測を繰り返せば
95%の確率で真値がこの範囲に入る

真値は一定で観測が確率的に変動する

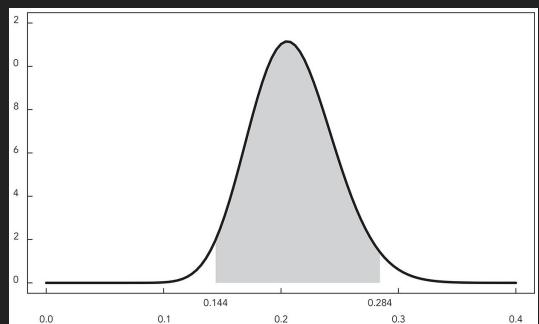
ベイズ法の発想

$P(\theta|D)$ を推定する



ベイズ法の発想

信用区間の意味



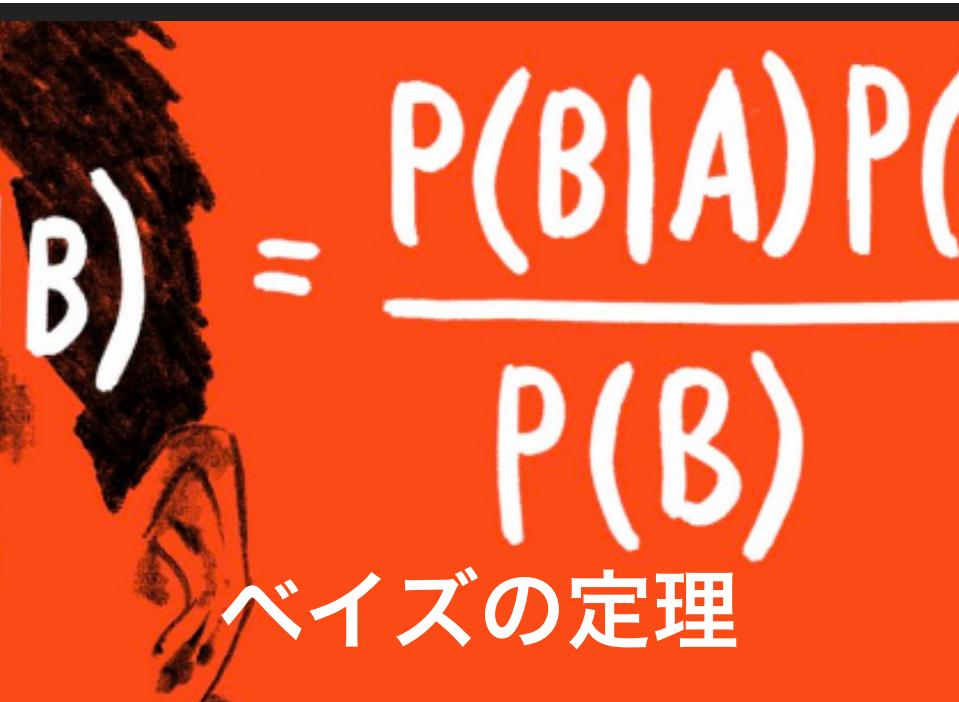
仮に真の値を仮定するなら
真値は95%の確率でこの範囲に入る

$$P(\theta|D)$$

はどう求めるのか？

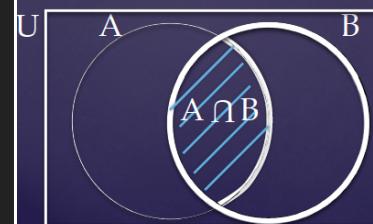
ここに

ベイズの定理あり



ベイズの定理

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$



P(A|B): 条件付き確率
事象Bのもとで、事象Aが起こる確率

P(B|A): 条件付き確率
事象Aのもとで、事象Bが起こる確率

P(A) or P(B): 周辺確率
事象Aのみ、事象Bのみが起こる確率

ベイズの定理 パラメータ推定に応用

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$A \rightarrow \theta$
 $B \rightarrow D$

事象Aを母数・パラメータ(θ)、
事象Bを得られたデータ(D)
に置き換える

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}$$

※ 离散変数の場合

ベイズの定理 パラメータ推定に応用

$P(\theta|D)$ データが得られた時に、母数が θ である確率

$P(D|\theta)$ 母数が θ である時に、データが得られる確率

$P(\theta)$ データによらず母数が発生する確率

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}$$

※ 离散変数の場合

ベイズの定理 パラメータ推定に応用

$P(\theta|D)$

事後確率

$P(D|\theta)$

尤度

$P(\theta)$

事前確率

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}$$

※ 离散変数の場合

ベイズの定理 パラメータ推定に応用

$P(\theta|D)$

事後確率

$P(D|\theta)$

尤度

$P(\theta)$

事前確率

$$\text{事後確率} = \frac{\text{尤度} \times \text{事前確率}}{\text{データが得られる確率}}$$

※ 离散変数の場合

ベイズの定理 パラメータ推定に応用

$\pi(\theta|D)$

事後分布

$f(D|\theta)$

尤度

$\pi(\theta)$

事前分布

$$\text{事後分布} = \frac{\text{尤度} \times \text{事前分布}}{\text{データが得られる確率}}$$

※ 連続変数の場合

ベイズの定理 パラメータ推定に応用

$f(D|\theta) \pi(\theta) \propto \pi(\theta|D)$

事後分布は尤度と事前分布の積に比例する

事前分布

prior

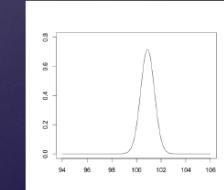
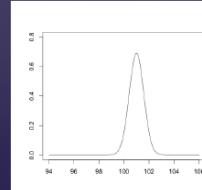
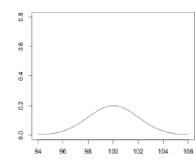
尤度

likelihood

\propto

事後分布

posterior



※分母のデータが得られる確率は、母数を含まないため比例定数になっている

事前分布...???

データに依存せずに θ が起こる確率...???

事前分布

- データ取得前の θ の分布に関する情報
- 研究者の仮説、先行研究、主観
- データから決める方法もある(客観ベイズ)

事前情報がない場合

事前分布
prior

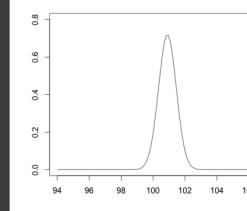
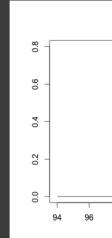
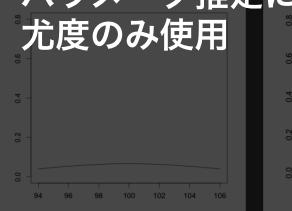


尤度
likelihood



事後分布
posterior

パラメータ推定に
尤度のみ使用



最尤推定と同様の推定結果になる
※推定するパラメータは分布する

ベイズ推定の特徴

パラメータの**分布**を推定

尤度に加えて**事前分布**を推定に使用する
要はこれだけ

MCMCによる ベイズ推定

MCMC

Marcov**Chain**Monte**Carlo** Methods

マルコフ連鎖モンテカルロ法

マルコフ連鎖モンテカルロ法

移動の方向が一つ前の状態に基づいて決まるルールを付与した 亂数シミュレーション によってパラメータの分布を生成する方法

MCMCの代表

メトロポリスヘイスティング法
ギブスサンプリングもその一種

なぜベイズ推定にMCMC??

- ・パラメータの分布を得るのに適してる
- ・解析的に解けないパラメータの分布を近似的に求められる

ギブスサンプリング

step1: x_1 の初期値を指定する

step2: x_2 を初期値で固定し, x_1 をサンプリング

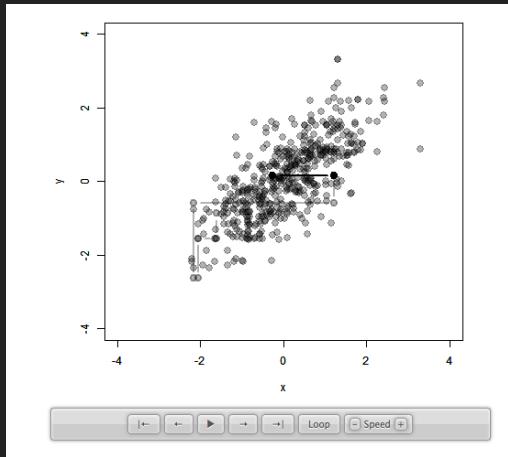
step3: x_1 をstep2の値に固定し, x_2 をサンプリング

step4: x_2 をstep3の値に固定して, x_1 をサンプリング

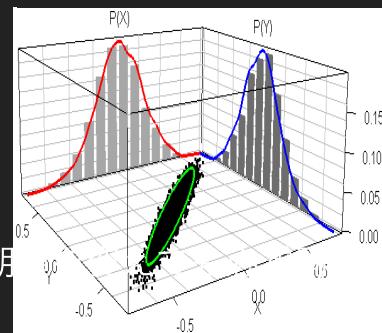
以下、 step3,4を繰り返す。

なるほどよくわからん

アニメーションで見てみましょう

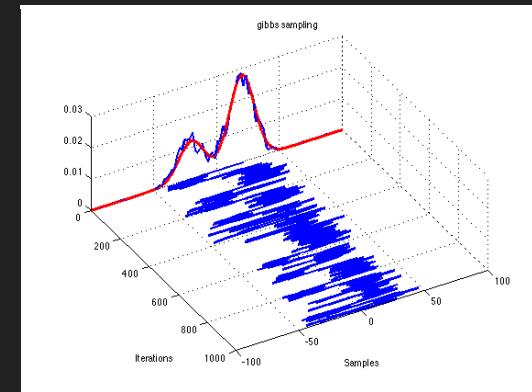


MCMCによるサンプリングが適切であれば
定常分布に収束する



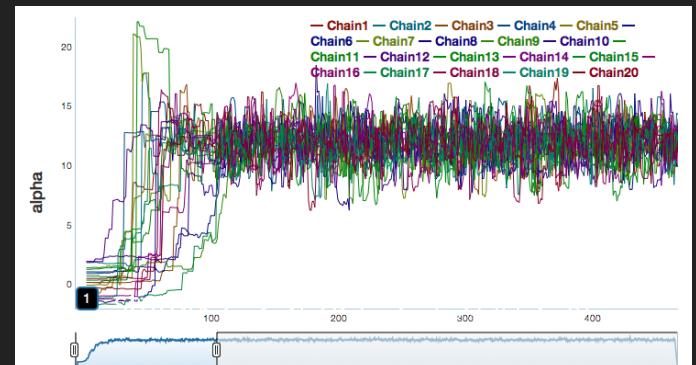
定常分布: 初期分布に達する

片方のパラメータで見ると…



BURNIN区間 WARM-UP期間

サンプリング初期は不安定なので推定に使わない

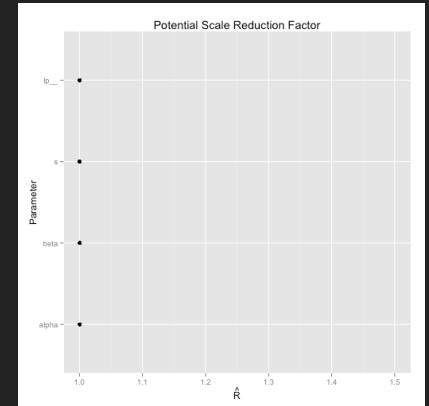


収束の診断指標

- GELMAN & RUBINの指標
- GEWEKEの指標
- 自己相関

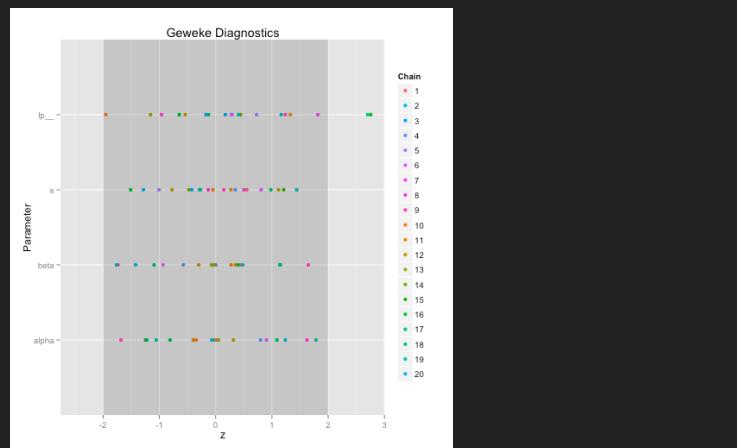
GELMAN & RUBINの指標

1.01を越えなければ良い
chain間の分散がchain内の分散より小さいと1に近づく
chain間の差を検討



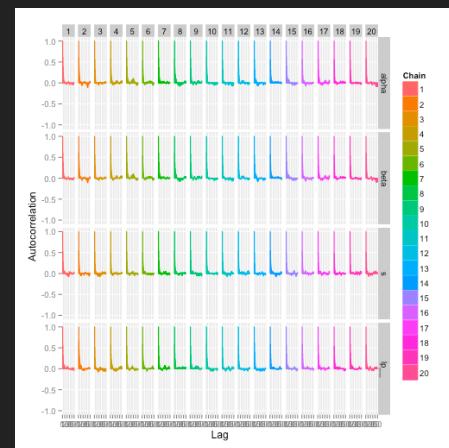
GEWEKEの指標

MCMCの最初の方と最後の方の平均に差がないか検定 Z 値が ± 1.96 以内に入っているれば差はない \rightarrow 収束している



自己相関

自己相関が5を超えると良くない



MCMCによるベイズ推定

MCMCによる推定の手順

- ①データを表現する統計モデルの記述
- ②パラメータの事前分布を設定
- ③MCMCの実行
- ④収束診断
- ⑤事後分布の解釈

回帰モデルの推定

- ①データを表現する統計モデルの記述

$$f(Y) = \alpha + \beta X + \varepsilon$$

Y は正規分布

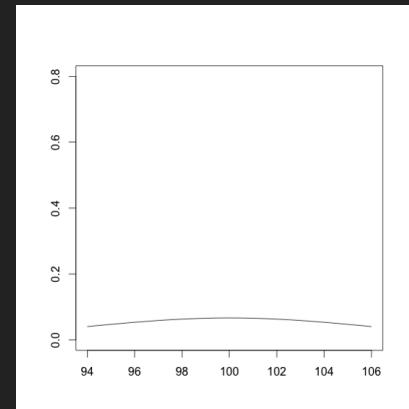
$$f(Y) = N(\alpha + \beta X, \varepsilon)$$

- ②パラメータ事前分布の設定

α と β の分布について情報なし

$$N(0, 100)$$

分散が大きい分布を指定



回帰モデルの推定

③MCMCのセッティング

- 収束回数
- chainの数
- burin区間
- thin(間引き数)

MCMCで回帰モデルの推定

$$f(Y) = \alpha + \beta X + \varepsilon$$

```
library(rstan)
N <- 500
x <- rnorm(N, mean = 50, sd = 10)
y <- 10 + 0.8 * x + rnorm(N, mean = 0, sd = 7)

stancode <- '
  data{ int<lower=0> N;
        real x[N];
        real y[N];
      }
  parameters { real alpha;
               real beta;
               real<lower=0> s;
             }
  model{ for(i in 1:N)
          y[i] ~ normal(alpha + beta * x[i], s); #推定するモデル
        }
```

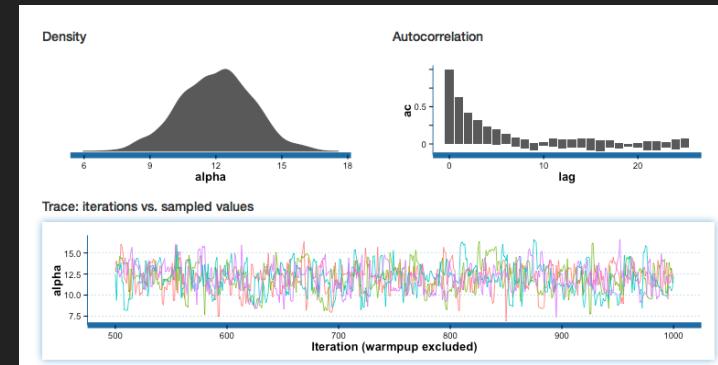
世界で二番目に簡単なStanコード

```
datastan <- list(N=N, x=x, y=y)
fit <- stan(model_code = stancode,data=datastan,iter=5000,chain=4)

my_shinystan <- as.shinystan(fit)
launch_shinystan(my_shinystan)
```

推定結果を見てみよう
On the shinyRStan

SHINYSTANの例



事後分布・自己相関
トレースプロット

事後分布の要約統計

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
alpha	1	527	12.0	1.6	8.7	12.1	15.2
beta	1	526	0.8	0.0	0.7	0.8	0.8
s	1	634	7.2	0.2	6.7	7.2	7.7
lp__	1	514	-1236.2	1.3	-1239.4	-1235.9	-1234.8

平均・標準偏差・信用区間

ベイズ推定が活ける
研究例

論文ではこんな風に報告

Model and parameter	EB posterior mean and percentiles			
	Mean	Median	P _{2.5}	P _{97.5}
Latent growth model for adolescent MRJ use and DLQ				
AL(1)	57.7	57.62	44.68	71.73
AL(2)	-32.86	-32.95	-47.05	-18.26
AL(3)	9.8	9.79	8.36	11.2
AL(4)	-1.57	-1.58	-2.97	-0.13
PS(1,1)	280.1	225.2	30.49	838
PS(2,2)	1102.8	912.2	108.9	3229

Ozecowski, T. J. (2014). Empirical bayes MCMC estimation for modeling treatment processes, mechanisms of change, and clinical outcome in small sample. *Journal of Consulting and Clinical Psychology*, doi: 10.1037/a0035889

8

ベイズ推定の報告の仕方

小症例数でのSEM

J Consult Clin Psychol. 2014 Oct;82(5):854-67. doi: 10.1037/a0035889. Epub 2014 Feb 10.

Empirical Bayes MCMC estimation for modeling treatment processes, mechanisms of change, and clinical outcomes in small samples.

Ozecowski TJ¹.

Author information

Abstract

OBJECTIVE: The current analysis demonstrates the use of empirical Bayes (EB) estimation methods with data-derived prior parameters for studying clinically intricate process-mechanism-outcome linkages using structural equation modeling (SEM) with small samples.

METHOD: The data were obtained from a small subsample of 23 families receiving Functional Family Therapy (FFT) for adolescent substance abuse during a completed randomized clinical trial. Two or 3 video-recorded FFT sessions were randomly selected for each family. The middle 20-min portion of each session was observed and coded. An SEM examining the influence of a select set of observed therapist behaviors on pre- to posttreatment change in mother reports of family functioning and, in turn, pre- to posttreatment change in adolescent reports of adolescent marijuana use and delinquent behavior was specified. The SEM was implemented using EB estimation with data-derived maximum likelihood (ML) prior parameters and Markov Chain Monte Carlo (MCMC) estimation of the joint posterior distribution.

RESULTS: The EB SEM results indicated that a relatively high proportion of individually focused general interventions (i.e., seek information, acknowledge) as well as relationally focused meaning change interventions by therapists during sessions of FFT were predictive of pre- to posttreatment increases in levels of family functioning as reported by mothers in families of substance-abusing adolescents. In turn, increases in mother-reported family functioning were predictive of reductions in levels of adolescent-reported delinquent behavior.

CONCLUSIONS: EB MCMC methods produced more stable results than did ML, especially regarding the variances on the change factors in the SEM. EB MCMC estimation is a viable alternative to ML estimation of SEMs in clinical research with prohibitively small samples.

マルチレベル因子分析



まとめ

- ベイズ推定はパラメータの分布を推定
- ベイズ推定は事前分布を利用する
- ベイズ推定はMCMCで行う
- RHAT等の指標で収束判断
- 複雑なモデルもMCMCがあれば怖くない

ENJOY!!