データ解析のための統計モデリング入門

8. マルコフ連鎖モンテカルロ(MCMC)法とベイズ統計モデル

河野和平

第7章一般化線形混合モデル

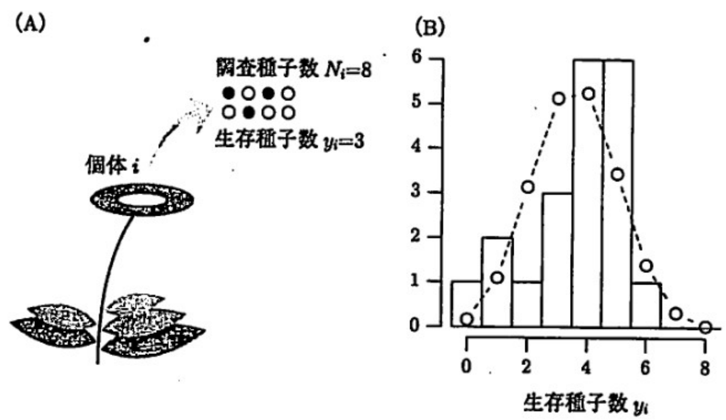
- ランダムに決定されるのは個体差 r_i のみ
  - r_i について積分すればよい
- K個のパラメータがある場合, K回の多重積分
  - 計算時間が長くなる
  - 最尤推定値の探索が困難

⇨ マルコフ連鎖モンテカルロ法の導入

例題:種子の生存確率(個体差なし)

- 植物 i (20個体)の生存種子数 y_i (MAX 8)を調査

$$\{y_1, y_2, \ldots, y_{20}\} = \{4,3,4,5,5,2,3,1,4,0,1,5,5,6,5,4,4,5,3,4\}$$



- 種子生存確率 q
- ある個体 i の生存種子数 y_iである確率

$$p(y_i|q) = \binom{8}{y_i} q^{y_i}(1-q)^{8-y_i}$$

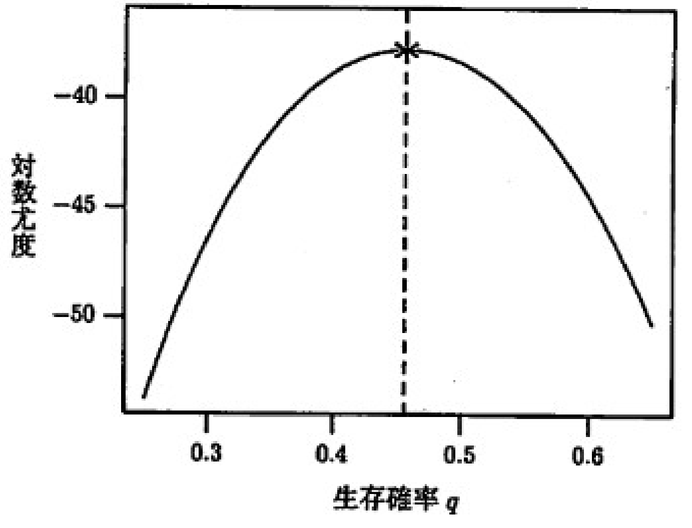
尤度L(q)

- 全個体における種子数y_iである確率の積

$$L(q)=p(Y|q)=\prod_i p(y_i|q)$$

$$\log L(q)=\sum_i y_i \log q+(8-y_i)\log(1-q)+\text{定数}$$

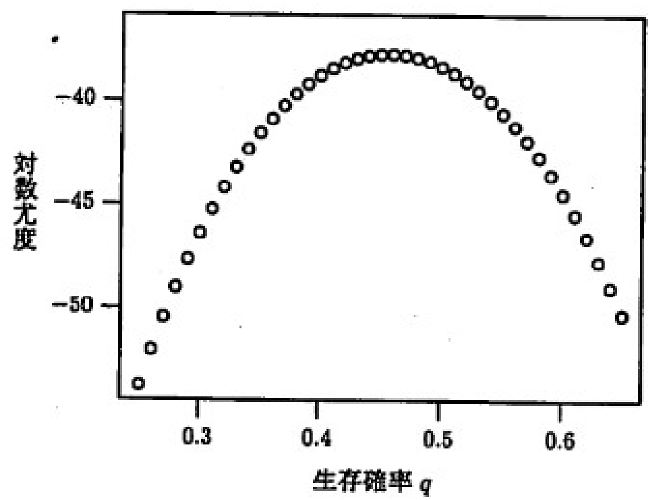
$$\hat{q}=0.45625$$



解析的に求められるとは限らない

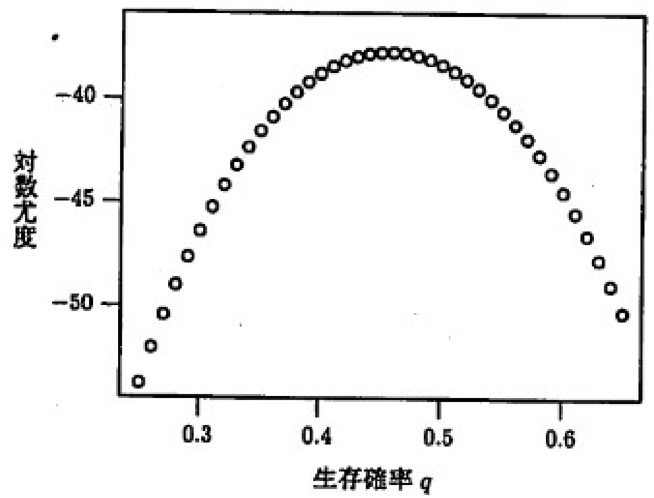
ふらふら試行錯誤(1)

- 簡単のため、離散値に変換
  - qは0.01きざみに値を取る



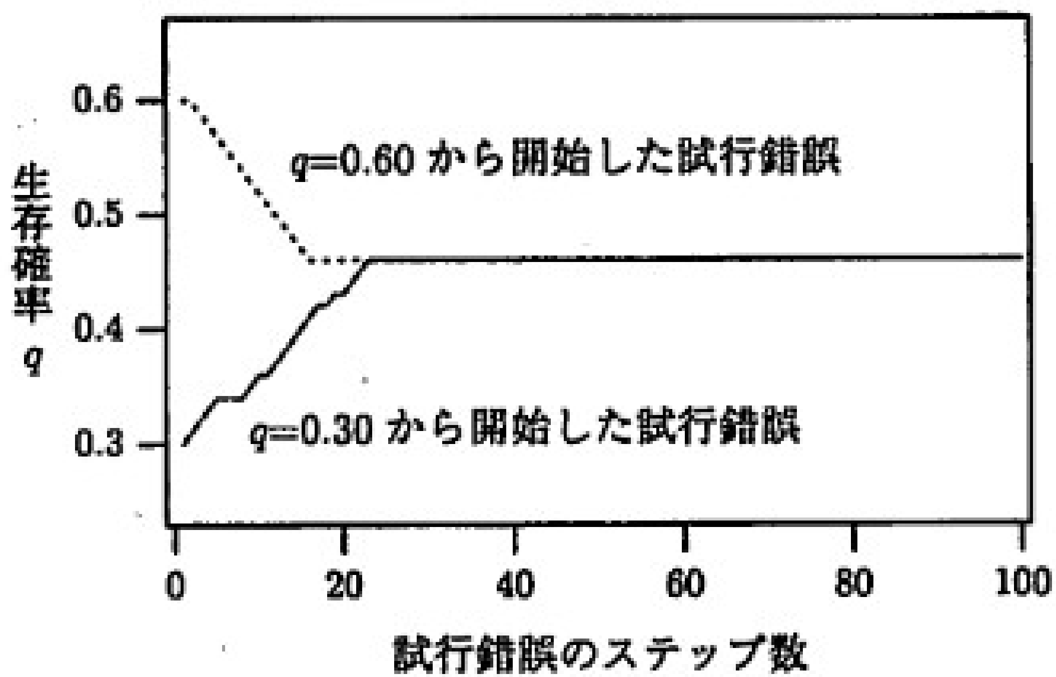
- qの初期値を与える
  - q=0.30 のとき対数尤度logL(q) = -46.38

ふらふら試行錯誤(2)



- 以下の処理を繰り返し行う。
  ①となりの2つからランダムに選ぶ
  ②対数尤度が現在値より大きければ移動

生存確率q



生存確率 q
0.6
0.5
0.4
0.3
q=0.60 から開始した試行錯誤
q=0.30 から開始した試行錯誤
0
20
40
60
80
100
試行錯誤のステップ数

メトロポリス法

- MCMC法のひとつで、ふらふら試行錯誤を改造
- アルゴリズム

  ①パラメータ q の初期値を決定

  ② q を増やすか減らすかをランダムに決定(q^{new})

  ③ $\log L(q) \leq \log L(q^{new})$ なら q を q^{new} に更新

  ④ $\log L(q) > \log L(q^{new})$ なら確率 r で q を q^{new} に更新

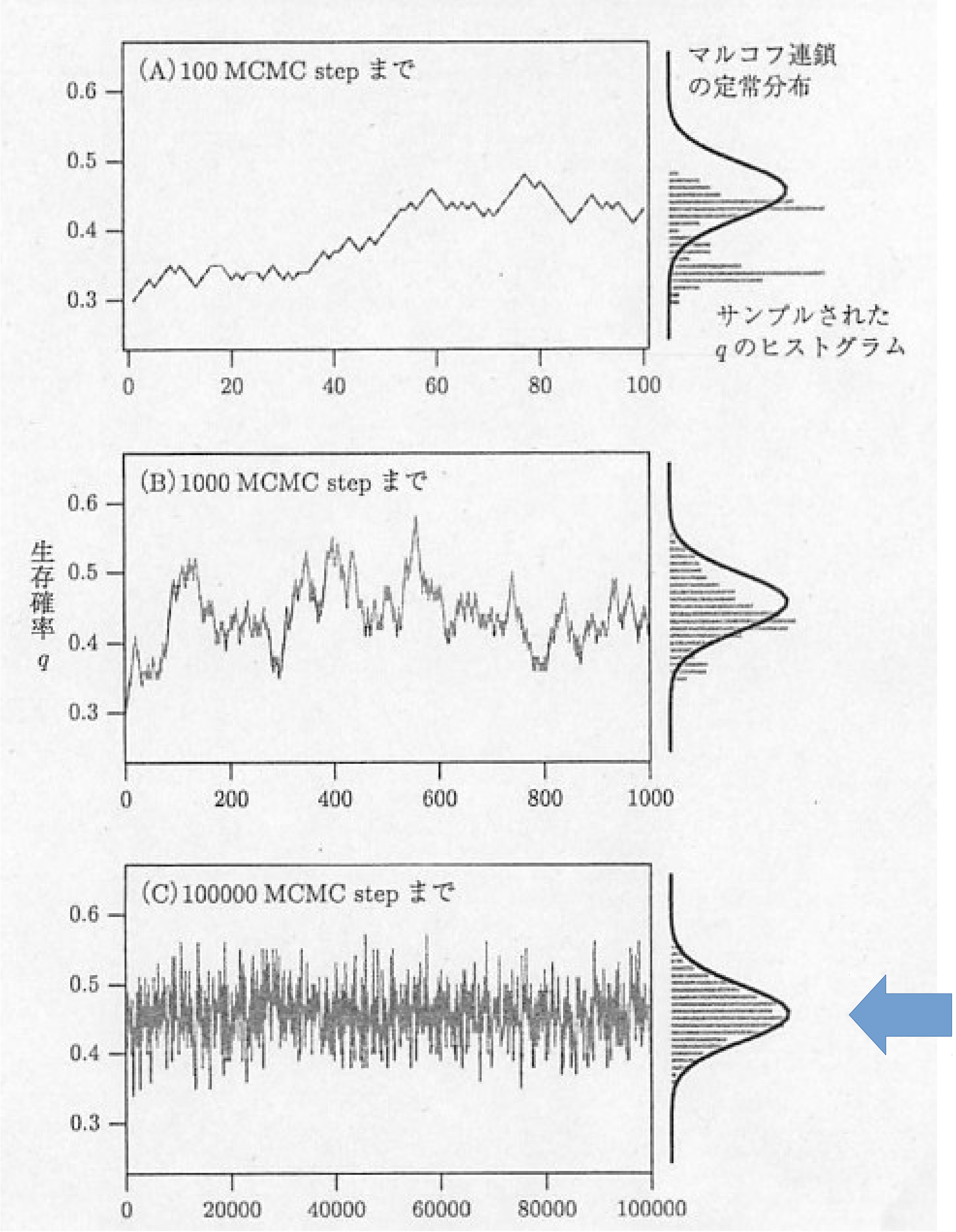
$$r = \frac{L(q^{new})}{L(q)}$$

$$\log L(q = 0.30) = -46.38$$

$$\log L(q^{new} = 0.29) = -47.62$$

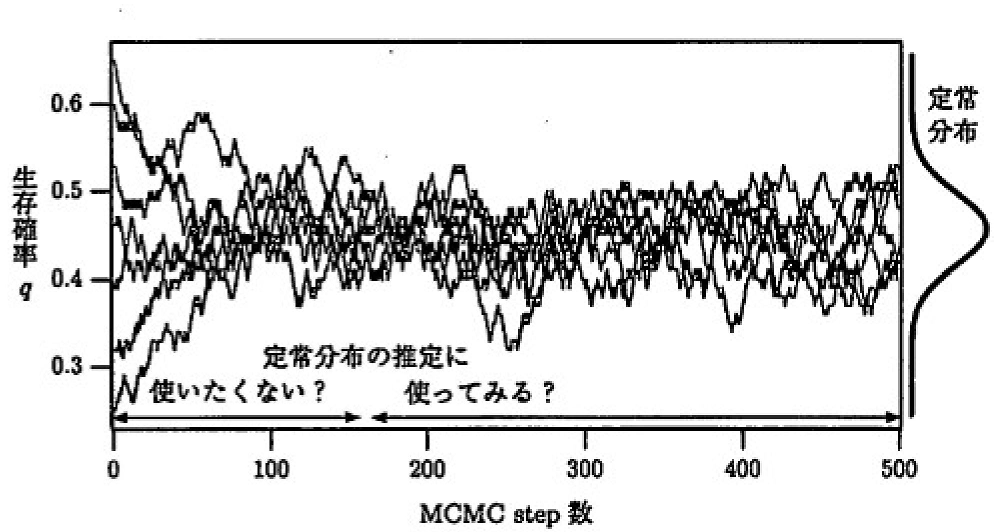
$$r = \exp(-47.62 + 46.38) = 0.29$$

生存確率q


(A) 100 MCMC step まで
(B) 1000 MCMC step まで
(C) 100000 MCMC step まで
マルコフ連鎖の定常分布
サンプルされたqのヒストグラム
生存確率 q
0.6
0.5
0.4
0.3
0
20
40
60
80
100
200
400
600
800
1000
20000
40000
60000
80000
100000

定常分布

複数の初期値q



どのような初期値qを与えてもStep数を重ねると定常分布に従う

定常分布

$$p(q|Y)=\frac{L(q)}{\sum_q L(q)}$$

$\sum_q L(q)$ ：データYにのみ依存する定数

$$p(q|Y)\propto L(q)$$

十分に長いStep数を重ねたメトロポリス法によって得られた
MCMCサンプルは定常分布 $p(q|Y)$ からのランダムサンプリングと言える。

ベイズの公式

尤度

$$p(Y|q)=L(q)$$

事前分布

事後分布

$$p(q|Y)=\frac{p(Y|q)\,p(q)}{\sum_q p(Y|q)\,p(q)}$$

$$\sum_q p(Y|q)\,p(q)=p(Y)$$

$$p(q|Y)=\frac{p(Y|q)\,p(q)}{p(Y)}\propto p(Y|q)\,p(q)$$

比較

- 数値実験の結果(定常分布)

$$p(q|Y)=\frac{L(q)}{\sum_q L(q)}=\frac{L(q)q(p)}{\sum_q L(q)q(p)}$$

- ベイズの公式

$$p(q|Y)=\frac{p(Y|q)\,p(q)}{p(Y)}\propto p(Y|q)\,p(q)$$

p(q)=定数（一様分布）なら2つの式は等価

二項分布の積である尤度L(q)とそのパラメータqの事前分布p(q)積に比例