HOKKAIDO UNIVERSITY

Title	最近のベイズ理論の進展と応用［I］ 階層ベイズモデルの基礎
Author(s)	久保, 拓弥
Citation	電子情報通信学会誌, 92(10), 881-885
Issue Date	2009-10-01
Doc URL	http://hdl.handle.net/2115/39717
Rights	copyright©2009 IEICE
Type	article
Note	講座
File Information	kubo2009IEICE.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

最近のベイズ理論の進展と応用［Ⅰ］
――階層ベイズモデルの基礎――

Recent Advances and Applications on Bayesian Theory［Ⅰ］:
An Introduction to Hierarchical Bayesian Models

久保拓弥

1. はじめに

階層ベイズモデルは従来の統計モデルでは扱いの難しかった様々な要因，例えば原因不明の個体差・地域差あるいは時間相関・空間相関の効果を扱うときに有用な考え方である[(1),(2)]．近年の計算機ハードウェア・ソフトウェアの進歩のおかげで，筆者のように複雑な推定プログラムを書けない研究者でも比較的容易に階層ベイズモデルを扱えるようになった．今回の解説では，そのような便利なソフトウェアを利用するときに必要となる，階層ベイズモデルの基本的な考え方を説明したい．

2. 架空例で理解する階層ベイズモデル

ここでは架空野球データ（表1）を使ったモデリング例を示しながら，「階層ベイズモデルとは何か？」を説明してみたい．

打者の打数・安打数（表1）から打率を推定する方法を検討してみよう．この架空のデータは以下のような性質を持つものとする．

・ 各チームの3～5番打者だけを選んだ

目　　次

［Ⅰ］ 階層ベイズモデルの基礎 （10月号）
［Ⅱ］ 逐次ベイズとデータ同化 （12月号）
［Ⅲ］ ノンパラメトリックベイズ （1月号）
［Ⅳ］ 変分ベイズ法 （2月号）
［Ⅴ・完］ モンテカルロ法の展開 （3月号）

久保拓弥　北海道大学地球環境科学研究院環境生物科学部門専攻
E-mail　kubo@ees.hokudai.ac.jp
Takuya KUBO, Nonmember (Graduate School of Environmental Earth Science, Hokkaido University, Sapporo-shi, 060-0810 Japan).
電子情報通信学会誌　Vol.92 No.10 pp.881-885 2009年10月
©電子情報通信学会 2009

表1　架空データ：20打者各 N_i 打席の安打数（Y_i）と打率（q_i）

i	N_i	Y_i	$\hat{q}_i$	真の q_i	i	N_i	Y_i	$\hat{q}_i$	真の q_i
1	43	7	0.163	0.273	11	46	11	0.239	0.301
2	42	11	0.262	0.280	12	41	11	0.268	0.303
3	42	6	0.143	0.284	13	49	17	0.347	0.305
4	50	14	0.280	0.287	14	46	16	0.348	0.307
5	44	10	0.227	0.289	15	41	14	0.341	0.309
6	42	15	0.357	0.291	16	46	14	0.304	0.311
7	44	14	0.318	0.293	17	49	15	0.306	0.314
8	44	13	0.295	0.295	18	49	16	0.327	0.317
9	45	10	0.222	0.297	19	48	19	0.396	0.321
10	49	8	0.163	0.299	20	44	15	0.341	0.328

・ 各打者は i という記号で表され，$i=1,2,3,\cdots,20$，つまり20打者いる
・ 各打者は N_i 打数のうち Y_i 回安打した

このようなデータから各打者の打率をできるだけ正確に推定したい，としよう．

打率 q_i とはある打席で打者 i が安打する「確率」と定義する．これは架空データなので，「真の打率」なるものが分かっていて表1に示されている[(注1)]．このような「真の打率」を知らないときに，与えられた打数・安打数データだけからできるだけ正確な，あるいはできるだけひどくない打率の推定値 $\hat{q}_i$ を評価できるような統計モデルを作りたい．このときに階層ベイズモデルの考え方が有用になり得る，ということを示してみよう．

2.1 統計モデルの部品：二項分布

打率を推定する統計モデルの部品として使う確率分布としては，二項分布が適切だろう．ある打率 q_i と打数 N_i のもとでの安打数を表現できる．例えば，表1の打者 $i=1$ の $N_1=43$ 打数の中で安打数 $Y_1=7$ 本となる

（注1） なお，この架空データは与えた「真の打率」に基づいて二項乱数で生成したものである．

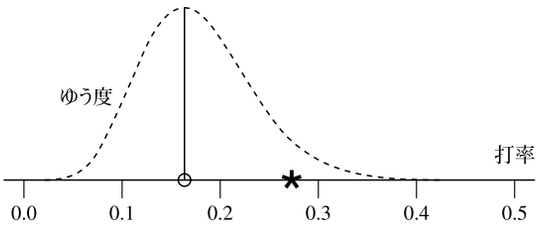


図1　43打数中7安打の打者 $i=1$ の対数ゆう度 log $L(q_1|Y_1)$（破線）　○は対数ゆう度が最大になる最ゆう推定値（7/43＝0.163），＊は打者 $i=1$ の「真の打率」（0.273，表1）である．

確率は二項分布の確率密度関数

$$f(Y_1=7\,|\,q_1)=\binom{43}{7}q_1^{\,7}(1-q_1)^{43-7}$$

で計算できる．このような確率密度関数 f（安打数 $|\,q_i$）を打率 q_i の関数とみなして $L(q_1|\,Y_1)=f(Y_1|\,q_1)$ としたものを尤度（以下，ゆう度）である．打率 q_1 に対するゆう度，つまり観測データに対する統計モデルの「当てはまりの良さ」を図示してみよう（図1）．

ゆう度 $L(q_1|\,Y_1)$ を最大化するような打率 q_1 が最ゆう（最ももっともらしい）推定値である．対数ゆう度 log $L(q_1|\,Y_1)$ を q_1 で微分して極値を求めると，最ゆう推定量は $\hat{q}_1=Y_1/N_1=7/43=0.163$ すなわち安打数／打数であることが分かる．

同じように全20打者の安打数が $\{Y_i\}=\{Y_1, Y_2, \cdots, Y_{20}\}$ であると観察される確率は各打者の $f(Y_i|\,q_i)$ を20打者分掛け合わせたものになり，全体データのゆう度関数は，

$$L(\{q_i\}\,|\,\{Y_i\})=\prod_{i=1}^{20}f(Y_i|\,q_i)$$

と定義される．各打者の打率の最ゆう推定量は「安打数（Y_i）/打数（N_i）」となる．これは表1に示している．

ところが，このようにして推定された打率は図2に示しているように，「真の打率」からは掛け離れたものとなった．

表1で示しているように，「真の打率」は0.273～0.328の範囲であり，どの打者も「おおよそ3割打者」といえる．しかしながら，推定された打率は0.143～0.396と広い範囲に散らばり，すなわち「チームの3～5番打者なのに2割も打てないやつ」から「4割打者」まで混在していることになってしまった．

もちろんこのように推定されてしまった原因は「データ不足」つまり打数 N_i が少なすぎるためである．それではこのデータは全くの無価値で，このめちゃくちゃな推定値はどうにも改良できないものなのだろうか？

2.2　ベイズモデル化：事前分布と事後分布

基本的にはデータのないところから情報は推定できない．したがって，打者 $i=1$ の「43打数中の7安打」というデータから真の打率0.273にかなり近い推定値は得られない．そこで個々の打者の打率の推定の改善はあきらめるにしても，全体の推定値の散らばりをもう少し小さくして，「真の打率」が分布している範囲に収めるようにしたい．

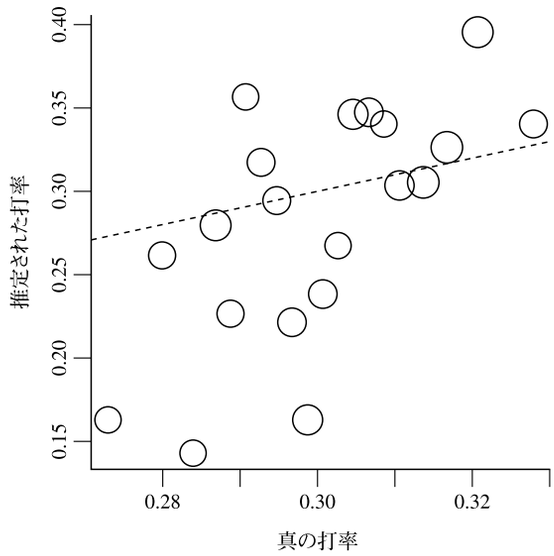


図2　割り算推定打率と真の打率　横軸は解析者が知り得ない「真の打率」，縦軸は安打数/打数で得られた推定打率．斜めの線は原点を通る傾き1の直線である．

そこで20名もの打者からデータを取ったことを利用して，「データの背後にある構造」を考慮した統計モデルを使って，個々の推定値が「めちゃくちゃ」にならないようにできないだろうか．

ここからの統計モデリングでは「打率の分布」を考える．つまり，そもそも各チームの3～5番打者なのだから打率はそんなに異なるものではない—どの程度「似ている」のかは現時点では分からないが—と仮定する．

ベイズモデルでは，推定したいパラメータが何か確率分布に従うと仮定し，この確率分布を事前分布と呼ぶ．ここでは打率 q_i を推定したいので事前分布は「各チームの3～5番打者全体」の打率の分布を表す．この事前分布を考慮しながら，各打者の安打数から推定される打率 q_i の確率分布が事後分布である．つまりベイズ推定では，打率の分布である事前分布を仮定し，ある打数のもとでの安打数という観測データから，事後分布という「打者 i の打率 q_i の確率分布」を推定する．

ここで「ベイズの公式」がこのような事前分布・観測データ・事後分布を関係付けるのに使われる．ベイズの公式とは

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

このような条件付き確率の関係を表すものである．この

例に沿って説明すると，ここで $p(P|D)$ はデータ（D，ここでは安打数）のもとでパラメータ（P，ここでは打率）が得られる確率である．$p(D|P)$ はその逆でパラメータ（打率）を決めたときにデータ（安打数）が得られる確率となるのでこれはゆう度である．$p(P)$ はあるパラメータ P が得られる確率なので事前分布，そして分母の $p(D)$ は「パラメータに関係なくデータ D が得られる確率」である．書き直してみると，

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

となる．分母の「データが得られる確率」は分子のゆう度×事前分布をすべての打率にわたって積分した値であり，定数なので以下では考えないことにする．

打率の事前分布と事後分布はどのような関係になっているのだろうか．もし何か「打率に関する適当に定めた事前分布」を決めることができたとしよう．図3で示している事前分布（灰色の線）は打率3割付近で高くなるような確率分布である．ここで打者 $i=1$ の打数43のうち安打数7といった観測データがあったとすると，事前分布・ゆう度・事後分布の関係から打者 $i=1$ の打率の事後分布（黒実線）が推定できる．

この「適当に定めた事前分布」から全20打者の事後分布を個々に推定すると図4のようになるだろう．

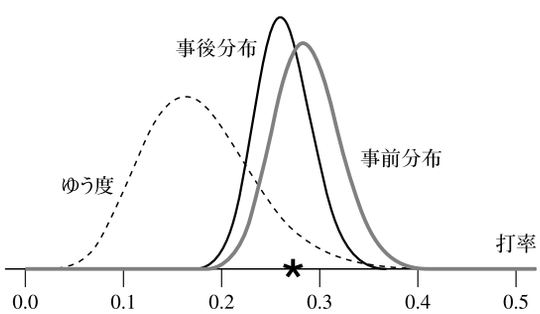


図3　打者 $i=1$ の打率 q_1 の事後分布　確率密度関数で事前分布・事後分布を図示する．図1と同じように破線で対数ゆう度を示している．灰色の線は適当に定めた事前分布，黒の実線は推定された事後分布，＊は打者 $i=1$ の「真の打率」(0.273，表1) である．

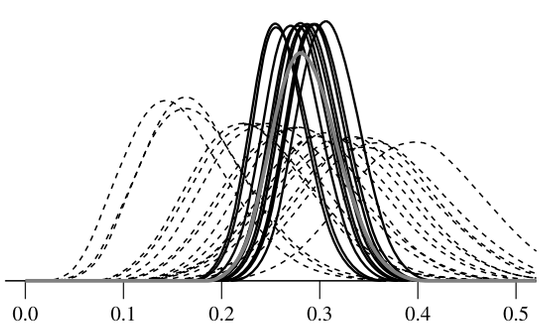


図4　全打者の打率 q_i の事後分布　全20打者について図3と同じように作図したもの．

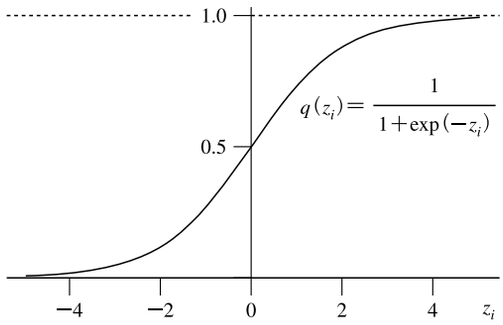


図5　ロジスティック曲線　z_i がどのような値をとっても $0<q(z_i)<1$ となる．

ここまでの話をまとめると，何か適切な事前分布つまり「打率の確率分布」を決めることができれば，ベイズモデルを使って全打者の打率の分布を考慮しつつ，個々の打者の打数・安打数からそれらしい事後分布が得られそうだ，ということになる．

残された問題は，どのようにして「適切な」打率の確率分布（事前分布）を決めればよいか，である．この問題を解決するためにベイズモデルを階層ベイズモデル化する必要がある．

2.3　打率を「全打者の平均」＋「打者差」に分割

この打率推定問題を階層ベイズモデル化するにはいろいろな方式がある．ここでは，モデル内で打率を「全体の平均」＋「打者差」と分割する定式化を試みてみよう．そのために安打する確率 $q(z_i)$ をロジスティック関数(概形は図5）で表現する．

ある打者 i の z_i を $z_i=a+b_i$ と置くことで，全打者の平均 a と打者差 b_i に分割できる．ここで a は全打者に共通の値であるが，b_i は個々の打者ごとに異なるものとする．どちらのパラメータも［$-\infty$，∞］の範囲でどのような値になってもよいのがこの定式化の利点の一つである．

2.4　打率の事前分布・事後分布の関係

先に述べたように「適切な打率の分布」を決める方法が分かればベイズモデルによって打率の事後分布が得られる．打率は今やロジスティック関数で逆変換された世界で $a+b_i$ と分割された．そこで打率そのものではなく，この全体の平均からのずれである，打者差 b_i の事前分布について考えることにしよう．

ここでは簡単のため，打者差 b_i の事前分布が平均ゼロで標準偏差 σ の正規分布で表現できるとしよう．

新しく導入されたパラメータ σ は「打者差のばらつきの大きさ」を表すものである．この σ が事前分布・事後分布に与える影響を「やや不正確な概念図」で示すと図6のようになる．

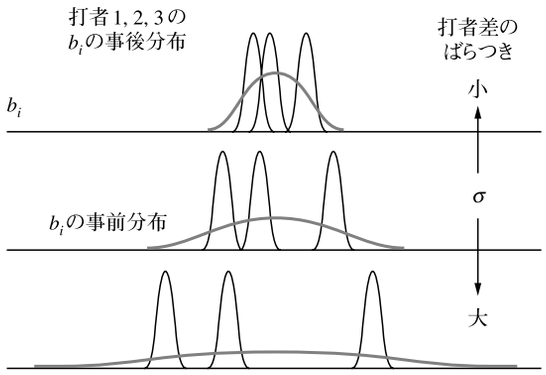


図6　打者差のばらつきと事後分布　事前分布の「幅」つまり「打者差 b_i のばらつきの大きさ」を決めるパラメータ σ の大小と b_i の事前分布（灰色線）並びに各打者（ここでは3打者だけ）の打者差 b_i の事後分布（黒線）の関係．なお，この図はやや不正確に描いている．

- σ がとても小さければ打者差 b_i の事後分布間の距離が狭まり，つまり「どの打者も打率がかなり似ている」ことになる
- σ がとても大きければ，b_i は各打者のゆう度関数に近づく，各打者のデータに引っ張られる

といった関係を柔軟に表現できる．

打者差 b_i の事前分布が σ すなわち「打者同士はお互いどれぐらい似てるか」を表すパラメータの大小が重要らしい，と分かった．それではこのパラメータはどのように定めればよいのか？

これは観測データに決めさせるのがよいだろう．打者差のばらつき σ のとる値が変わることで，打者差 b_i の事後分布が変化し観測データへの当てはまり具合が変わる．ところで，事後確率の密度関数の極大値の「高さ」を（ここでは仮に）事後確率と呼ぶことにする．あるパラメータのもとで事後確率が大きくなっていれば，それはデータへの当てはまりが良いことに対応している．

事後分布の定義から，事後確率の大小はゆう度と打者差 b_i の事前分布である（他の事前分布は無情報事前分布なので影響は少ない）．つまり σ が小さければ事後確率は高くなり，大きければ低くなる．

したがって，事後確率を大きくするような σ があって，それは個々の打者のデータに合うような b_i を生成し得る事前分布でありながら，同時にできるだけその「幅」が狭いもの，ということになる．

2.5　階層ベイズモデル化：超事前分布の利用

階層ベイズモデルでは，事前分布のパラメータである σ の更に事前分布，つまり超事前分布の導入によって解決する．事前分布のパラメータが更に別の事前分布を持つ，といった階層構造を持つので，このようなベイズモデルは階層ベイズモデルと呼ばれる．

打者差 b_i の散らばり具合は標準偏差 σ の事前分布によって制約されていたけれど，一方で打者差のばらつきを表すパラメータ σ そのものは1個しかないので，そのような制約を設けられないし，また σ に関する事前情報は何もないので0より大きければどんな値をとっても差し支えない．

そこで σ の事前分布を無情報事前分布[3]，すなわち $\sigma>0$ かつひたすらに「平べったい」確率分布とする．正規分布の分散（σ^2）パラメータの共役な事前分布は逆ガンマ分布なので，分散の逆数 τ（$=1/\sigma^2$）の事前分布がガンマ分布となる．このときに定数パラメータを設定して，分散の大変大きな確率分布となるようにする．

この階層ベイズモデルの中で打率は「全打者の平均」a と「打者差」b_i に分割されているので，全打者が共有するパラメータ a も推定しなければならない．このパラメータの事前分布も無情報事前分布とする．全打者の平均 a は［$-\infty$，∞］の範囲で自由な値をとっても構わないので，平均ゼロで標準偏差100の「平べったい」正規分布とした．

3.　MCMCによる階層ベイズモデルの推定

ここまでで，階層ベイズモデルの部品である「観測データとパラメータを対応付けるゆう度」「パラメータの事前分布」「事前分布のパラメータの超事前分布」がすべてそろった．このように設計された階層ベイズモデルのパラメータの事後分布を推定する方法としては経験ベイズ法[1],[2]とMarkov Chain Monte Carlo（MCMC）法がよく使われている[3]～[5]．ここではMCMC法で事後分布を推定してみよう．

階層ベイズモデルの事後分布はゆう度・事前分布・超事前分布の積に比例すると定義されるので，数値的に確率分布の評価は困難である．そこで事後分布を直接に計算するのではなく，MCMC計算によって事後分布からのランダムサンプルセットを得る．これによって各パラメータの事後分布の平均値・中央値・95%区間などが分かる．

公開されているソフトウェアWinBUGS[6]は簡単な操作で，MCMC計算による階層ベイズモデルを含むベイズモデル一般の事後分布からのサンプリングによく使われている．

表1の打数・安打数データに基づいて，打率に関する階層ベイズモデルをWinBUGSで評価させると（詳しい手順はこの記事のサポートWebページを参照[7]），全打者の平均 a，打者差 $\{b_1, b_2, \cdots, b_{20}\}$，打者差のばらつき σ の事後分布からのランダムサンプルが得られた．各パラメータのサンプル中央値を使って打者ごとの打率 q_i を推定して図示したものが図7である．打者ごとに安打数/打数を打率の推定値とした場合に比べて，図7で示されている階層ベイズモデルで推定された打率はより妥

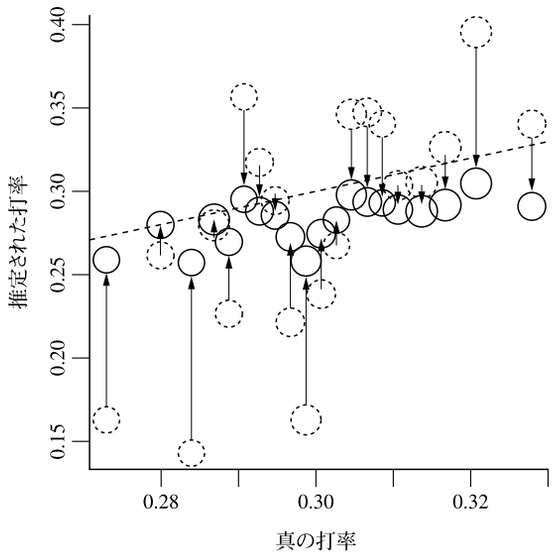


図7　階層ベイズモデルの推定結果　図2と同じで，横軸は「真の打率」，縦軸は推定された打率．黒ふち線の白丸が階層ベイズモデルによって推定された打率 q_i，灰色のふち線の白丸が各打者ごとに安打数/打数で推定された打率（図2）．矢印は階層ベイズモデルによる「補正」を表す．

当な範囲に分布しているようにみえる．

特に打者当りの打数が少なすぎるデータを使う場合には，個々の打者ごとに独立に割り算で打率を推定するのではなく，データの背後にある構造，例えば「全打者の打率は何か確率分布に従う」といったことを仮定すると推定は改善され得る，と考えてよいだろう．

4.　階層ベイズモデルの利点と応用

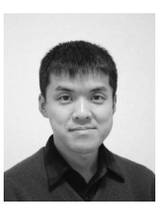
階層ベイズモデルが普及する以前には「主観事前分布」を使ったベイズモデルがよく使われていた．主観事前分布とは，この例題でいえば，例えば解析者の信念などに基づいて「各チームの3～5番打者の打率の分布は平均0.3で標準偏差0.02の事前分布に従う」と決めてしまう．推定計算が比較的容易になるが，事前分布の設定における恣意性が問題になる場合もある．

階層ベイズモデルはモデル中の重要な事前分布をいわば「データに決めさせる」ので，解析者の主観などに左右されにくい方式といえる．このために自然科学の諸分野での応用が広まっている．筆者は動物・植物の生態学データの解析を専門としているが，例えば生態学の分野においても，ウミガメ上陸数の長期時系列データの解析，森林三次元構造の推定，アリの行動観察データの解析に階層ベイズモデルを応用している．

文　　献

（1）石黒真木夫，松本　隆，乾　敏郎，田邉國士，階層ベイズモデルとその周辺，岩波書店，東京，2005.
（2）久保拓弥，“階層モデルで「個性」をとらえる，”数学セミナー，vol.46, no.11, pp.16-22, Nov. 2007.
（3）マルコフ連鎖モンテカルロ法，豊田秀樹（編著），朝倉書店，東京，2008.
（4）伊庭幸人，ベイズ統計と統計物理，岩波書店，東京，2003.
（5）伊庭幸人，種村正美，大森裕浩，和合　肇，佐藤整尚，高橋明彦，計算統計Ⅱマルコフ連鎖モンテカルロ法とその周辺，岩波書店，東京，2005.
（6）MRC Biostatistics Unit. The BUGS project. http://www.mrc-bsu.cam.ac.uk/bugs/
（7）久保拓弥，生態学のデータ解析―信学会誌ベイズ解説．http://hosho.ees.hokudai.ac.jp/~kubo/ce/IeiceBayes2009.html.

（平成21年3月30日受付）



久保（くぼ）　拓弥（たくや）
平5九大・理卒．平10同大学院博士課程了．現在，北海道大学地球環境科学研究院助教．様々な生態学データの統計モデリング，計算生態学が専門．