階層モデルで「個性」をとらえる

久保拓弥. 2007. 階層モデルで「個性」をとらえる.
数学セミナー 46 (11): 16-22 (2007 年 11 月号, 通巻 554)

久保拓弥
北海道大学大学院地球環境科学研究院

観察データと統計モデリング

科学では観察・実験で得られたデータ— 構造をもった数値・記号のあつまり— をあつかいます．このとき統計学的な手法をもちいて，観察データにみられるパターンを説明できるようなうまい統計モデルを構築します．これによってデータとモデルを組みあわせて，モデルを特徴づけるパラメータなどを推定します．

このような統計モデリングこそがデータ解析の本質なのですが，多くの科学研究者はデータの処理を創造的なモデリングだとは気づいていません．むしろ，何かお役所仕事みたいな，誰かに定められてしまった意味のよくわからぬ手つづきみたいなものだと考えているようです．

ここからは (筆者が専門としている) 生態学であつかうようなデータ例にそって「よくわかる」統計モデリングについて考えていきます．

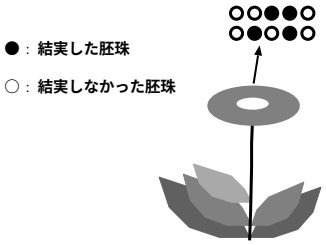

図 1 架空植物の 10 個の胚珠の結実調査

現実の生態学研究の事例は複雑すぎるので，図 1 のような架空植物のごく単純化した架空データを検討しましょう．この植物のある一個体を選んだときに，それが何個ぐらい種子を作るのか知りたい，とします．この植物は胚珠という種子のもとになる器管をどの個体も必ず 10 個もっています．つまり観察される種子数は最小 0 個で最大 10 個となります (図 1 の例では 4 個が結実しています)．胚珠が種子になることを結実，ある胚珠が種子になる確率のことを結実確率とよぶことにします．この結実確率の大小を決める生物学的な要因にはさまざまなものがあります．しかしながら，ここでは研究者はそれについて何もわかっていない，と仮定します．

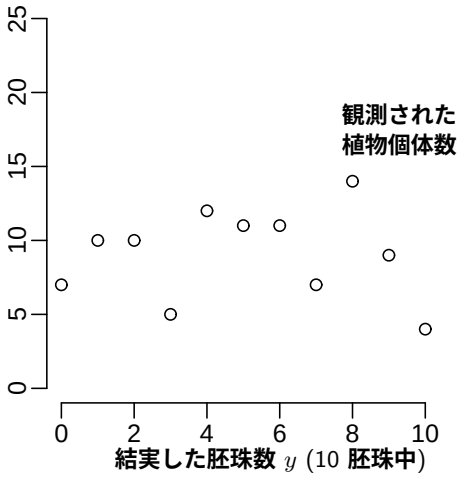

図 2 架空植物の観察結果: 度数分布

さてこの植物 100 個体を観察したときに，各個体の種子数について図 2 のようなデータが得られたとします．横軸は 10 胚珠中の結実した種子数 (y)，縦軸は頻度 (種子数 y だった個体数) になります (この観察データは `http://hosho.ees.hokudai.ac.jp/~kubo/ce/SuSemi2007.html` からダウンロードできます)．100 個体 × 10 胚珠から合計 496 個の結実した種子が得られました．

研究者たちはこのようなデータをどうあつかうのでしょうか．「ようするに 1000 個の全胚珠中 496 個が種子になったんだからこの植物の結実確率 0.496，個体あたりの平均結実種子数は 4.96 個，それでいいんでしょ? 標準偏差がないとケチつけられるからそれも計算して，あとは “ゆーい” かどうか “検定にかければ” いいんでしょ?」といった

何も考えていないルーチンワーク的な処理によって「データ解析，無事終了!」としてしまいたい研究者をよくみかけます.

いつもいつもこのように平均値その他の計算を「決まりきった手順」などと称して実施していれば，それで現象の観察データに見られるパターンは説明されたことになるのでしょうか?

割算推定量とその統計モデル

まずはこの場合の平均値計算 (結実胚珠数を全体の胚珠数で割ること) によって得られる確率の推定量は，どういう統計モデルにもとづくものなのか検討しましょう.

これは最尤 (さいゆう) 推定からでてくるものです．この架空植物のひとつの個体を i とよびます ($i = 1, 2, \cdots, 100$)．**すべての個体で結実確率 q が共通している**と仮定します．この仮定が正しいとすると，個体 i の 10 胚珠の中で結実した胚珠数が y_i 個となる確率は二項分布

$$f(y_i \mid q) = \binom{10}{y_i} q^{y_i}(1-q)^{10-y_i},$$

で表現できるということです．植物 100 個体の観察値 $\{y_i\} = \{y_1, y_2, \cdots, y_{100}\}$ が観察される確率は上の $f(y_i \mid q)$ を 100 個体ぶん掛けあわせたものになります．このときに，逆に観察データ $\{y_i\}$ が与えられたもので，パラメータ q は値が自由にとりうると考えると，この 100 個体ぶんの確率はパラメータ q の関数となります．これは尤度とよばれ，形式的には

$$L(q \mid \{y_i\}) = \prod_{i=1}^{100} f(y_i \mid q),$$

と定義されます．この尤度 $L(...)$ を最大化するパラメータの推定量 $\hat{p}$，を計算してみましょう．対数尤度をとって

$$\log L(q \mid \{y_i\}) = \sum_{i=1}^{100} \log \binom{10}{y_i} + \sum_{i=1}^{100} \Big\{ y_i \log(q) + (10 - y_i) \log(1-q) \Big\},$$

この q に関する微分がゼロになる q を計算すると，たしかに $\hat{q} = \sum_{i=1}^{100} y_i / 1000$，つまり結実した全胚珠個数を全胚珠個数で割った数になっています.

個体差を無視したモデルの予測

ところで，二項分布の最尤推定値 $\hat{q} = 0.496$ は図 2 に示されたような現象を説明しているのでしょうか? 観察データに統計モデルの予測，$100 f(y \mid \hat{q})$，を重ねてみると図 3 のようになります．これは表・裏の出現が同じ確率になるコイン投げを各人 10 回，合計 100 人にやってもらったときに，表がでた回数が y 回だった人数の分布とほぼ同じです.

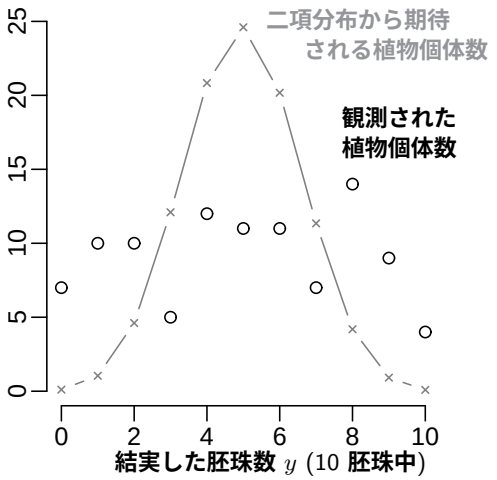

図 3 観察結果と二項分布モデル

結実確率 0.496 の二項分布モデルの予測と観察データを比較すると，次のようなちがいに気づきます:

- 10 個中 5 胚珠が結実する個体数は 24.6 と期待されるが観察データでは 11 個体しかいない
- 結実数 0 個の期待個体数は 0.11 なのに 7 個体，結実数 10 個の期待個体数は 0.09 なのに 4 個体が出現した

これを見ると，結実する確率 q は平均値計算で推定してしまえばよい，とする二項分布モデルでは現象をうまく説明できていない，と見当がつきます．おそらく,「どの個体でも胚珠が結実する確率 q は同じ (この例だと 0.5 ぐらい?)」という二項分布の仮定があまり正しくないのでしょう．このように，個体 i の結実種子数 y_i のばらつきが二項分布モデルの予測から逸脱してしまう現象を過分散 (overdispersion) とよびます.

図 2 に示されている観察データのパターンを説明するためには，二項分布モデルを拡張しなければなりません．たとえば，**結実する確率 q は植物個体によって異なる**らしいと考えてみるのは自然なことです．個体ごとに結実確率が集団平均からずれていることを，仮に個性もしくは個体差とよぶことにします．なぜ個体に差が生じるのだろうか? といった生物学的な問題は最後に議論することにして，ここでは目の前のデータに見え隠れしている個体差のあつかいに集中することにしましょう．

個体差を考慮した統計モデル

結実する確率 q が個体によって異なるよう統計モデルを拡張する準備として，結実する確率 q をロジスティック関数 $q(z) = \frac{1}{1+\exp(-z)}$ で表現することにします (図 4).

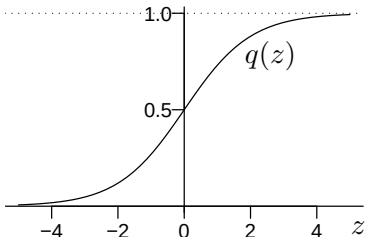

図 4 ロジスティック曲線 $q(z)$

次に，ある個体 i の z を z_i として，$z_i = \beta + \alpha_i$ となるように全個体共通の部分 β と個体差 α_i の部分に分割します.

さて，このように個体差をあらわすパラメータ α_i で改良した統計モデルを使って，観察データからパラメータを推定するにはどうしたらよいのでしょうか? 個性を考慮しない二項分布モデルの尤度方程式の q を $q(\beta + \alpha_i)$ に置きかえてみると，

$$L(\beta, \{\alpha_i\} \mid \{y_i\}) = \prod_{i=1}^{100} f(y_i \mid q(\beta + \alpha_i)),$$

となります．個体差なしモデルのときのように，この尤度を最大化するようなパラメータを推定すればよいのでしょうか? 個体差をあらわすパラメータ $\{\alpha_1, \alpha_2, \cdots, \alpha_{100}\}$ は 100 個あり，これに加えて結実する確率のうち全個体共通部分 β を加えると 101 個のパラメータを推定することになります．多少工夫すれば推定すべきパラメータ数を 100 個にできます．いずれにせよ 100 個体の挙動を説明するために 100 個以上のパラメータの推定値を確定しています.

これは「個体 i が結実する確率は $y_i/10$」，つまり結実する確率は個体ごとにいちいち割算推定量で計算する，という方式です．これでは「この植物はどのように結実するか」に関する説明ができたような気分になりません．いっそのこと，個体差なしモデルによる「どの個体も結実する確率は 0.496」といった大雑把な推定のほうがマシに思えます.

階層ベイズモデルで表現する個体差

個体差なしモデルでは観察データにみられるパターンがうまく説明できてるようには見えない (図 3)，しかし個体差パラメータ $\{\alpha_i\}$ を 100 個も最尤推定するのはいかにもおかしい，という状況を改善するために階層ベイズモデル (あるいは階層モデル) を導入します.

この例題におけるベイズモデルは何なのか，を端的に言うと全 100 個体の個体差 α_i をいちいち最尤推定しない，ということになります (このベイズモデルは伊庭氏の解説でのベイジアンモデリングと同じ，そちらの解説も参照してください). たとえば個体番号 $i = 1$ の架空植物の個体差 α_1 が $-1.2345\cdots$ などと確定できるはずだとは考えないで，α_1 は -3 ぐらいかもしれないし $+0.5$ ぐらいかも，などといいかげんなまま放置する，つまり α_i それぞれを確率変数で表現することにします.

しかしながら，個体差 α_i の確率分布は好き勝手に決めてよし，と許可してしまうとかなり無秩序な推定結果になります．そこで，各 α_i の確率分布は観察データ $\{y_i\}$ と「観察された 100 個体の結実確率には，どこか似ている部分がある」というルールによって制約してしまいたい，つまり「観察データをうまく説明できる範囲で，個体たちはできるだけ似ている (α_i がゼロに近い) となるように α_i を決めようね」と，なかなか都合の

よいことをもくろんでいるわけです．このように $\{\alpha_i\}$ を制約する役目を与えられた確率分布をベイズ統計学では事前分布とよびます．これに対して，観察データと事前分布で決まる α_i の確率分布は事後分布です．

この個体差 α_i の事前分布は，ここでは簡単のため平均ゼロで標準偏差 σ の正規分布

$$g_\alpha(\alpha_i \mid \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-\alpha_i^2}{2\sigma^2},$$

で表現できることにしましょう．観察された個体全体に共通するパラメータ σ は，この植物の個体たちはおたがいどれぐらい似ているかをあらわしていて，たとえば，

- σ がとても小さければ個体差 α_i はどれもゼロちかくになりますから「どの個体もおたがい似ている」
- σ がとても大きければ，α_i は各個体の結実数 y_i にあわせるような値をとる

といった状況が表現できるようになりました．ある個体 i の α_i の事後分布が事前分布 $g_\alpha(\alpha_i \mid \sigma)$ に依存している様子を図 5 に示します．

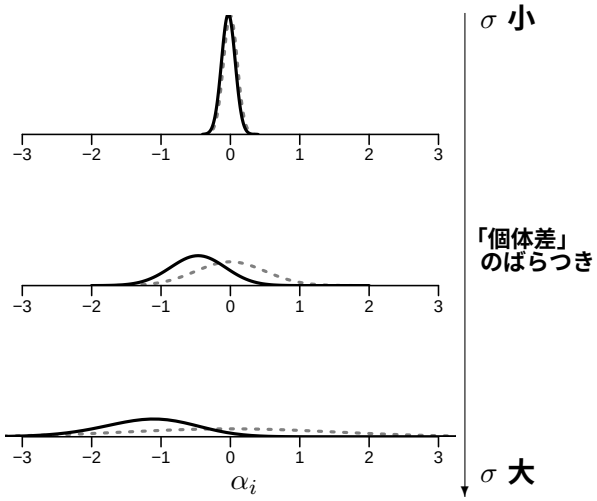

図 5 事前分布に依存する事後分布

ここで灰色の破線は $\{\alpha_i\}$ の事前分布，黒い実線はある個体 i の α_i の事後分布 (10 胚珠中の結実数 $y_i = 2$) です．全体のばらつき σ が小さいと α_i はゼロに近く，σ が大きいときには事前分布による制約が弱くなるのでゼロからずれています．

観察された 100 個体の集団で個体差のばらつきをあらわすパラメータ σ をどう決めるか，が重要な問題になります．しかし，ここではこの問題を先おくりすることにして，ただ単にパラメータ σ もまた何か確率分布 $h(\sigma)$ にしたがう確率変数だ，ということにしてしまいます．これは事前分布のパラメータの事前分布なので超事前分布とよばれています．そしてこんなふうにパラメータを何もかも確率変数にするのであれば，は全個体共通部分をあらわすパラメータ β も確率分布 $g_\beta(\beta)$ にしたがうとしましょう．これってどういう確率分布なの，といった疑問なんかもあとまわしにします．

めんどうな問題をすべて先おくりにしたまま，ベイズの公式にもとづいて観察データ $\{y_i\}$ のもとでのパラメータの同時分布は

$$p(\beta, \{\alpha_i\}, \sigma \mid \{y_i\}) = \frac{\prod_{i=1}^{100} f(y_i \mid q(\beta + \alpha_i))\ g_\beta(\beta)\ g_\alpha(\alpha_i \mid \sigma)\ h(\sigma)}{\int\int\int (\text{分子} \uparrow \text{そのまま})\ d\alpha_i\ d\sigma\ d\beta}$$

となり，この分母は定数なので，分子だけに注目すると

$$p(\beta, \{\alpha_i\}, \sigma \mid \{y_i\}) \propto \prod_{i=1}^{100} f(y_i \mid q(\beta + \alpha_i))\ g_\beta(\beta)\ g_\alpha(\alpha_i \mid \sigma)\ h(\sigma),$$

すなわち，事後分布の確率密度は尤度 (観察データのもとでの) と事前分布・超事前分布の確率密度の積になっています．個体差パラメータ α_i の事後分布を得るためにその事前分布 $g_\alpha(\alpha_i \mid \sigma)$ が必要で，さらにこの事前分布を決めるために超事前分布 $h(\sigma)$ が必要になる，といった階層構造があるので，このような統計モデルは階層ベイズモデルと呼ばれています [1]．この階層ベイズモデルを使うことで，この架空植物の結実確率に関する説明は改善されるのでしょうか?

経験ベイズ法による最尤推定

階層ベイズモデルのパラメータを推定する方法としては**経験ベイズ法**と **Markov Chain Monte Carlo (MCMC) 法**がよく使われています．

まず経験ベイズ法について説明しましょう．前の節で定義した事後分布 $p(\beta, \{\alpha_i\}, \sigma \mid \{y_i\})$ において全個体共通部分のパラメータ β の事前分布 $g_\beta(\beta)$ と個体差のばらつきをあらわす σ の (超) 事前分布 $h(\sigma)$ を「分散がとても大きな一様分布」にしてしまいます．これは言いかえると「β も σ も (観察データにあうように) 好き勝手な値をとっていいよ (ただし σ は標準偏差なので $\sigma > 0$)」と設定していることになります．いっぽうで，図 5 のように，各個体の個体差 α_i は平均ゼロかつ標準偏差 σ の正規分布である事前分布 $g_\alpha(\alpha_i \mid \sigma)$ に制約されているので，好き勝手な値をとることはできません．

一様分布の仮定によって $g_\beta(\beta)$ と $h(\sigma)$ が何やら都合よく「定数みたいなもの」に変えらてしまったので，事後分布は

$$\prod_{i=1}^{100} f(y_i \mid q(\beta + \alpha_i))\ g_\alpha(\alpha_i \mid \sigma),$$

に比例する量となり，さらにこの α_i について積分した量は

$$L(\beta, \sigma \mid \{y_i\}) = \prod_{i=1}^{100} \int_{-\infty}^{\infty} f(y_i \mid q(\beta + \alpha_i))\ g_\alpha(\alpha_i, \sigma) d\alpha_i \times (\text{定数}),$$

というふうに，観察データ $\{y_i\}$ のもとでのパラメータ β と σ の尤度方程式とみなせます．

あとは尤度 $L(\beta, \sigma \mid \{y_i\})$ を最大化するような $\hat{\beta}$ と $\hat{\sigma}$ を推定すればよいだけです．この推定計算は数値的な最尤推定法となり，そのプログラミングは多少めんどうなものとなります．

この数値計算を簡単にすませる抜け道があります．上のような尤度であらわされる統計モデルは一般化線形混合モデル (generalized linear mixed model あるいは GLMM) というクラスのモデルとまったく同じ形式になります [2]．この種子結実の統計モデルのような (個体差が $\{\alpha_i\}$ だけであるような) 単純な GLMM の場合，統計ソフトウェア R を使うことで簡単に推定計算できます [3]．R に `glmmML` というライブラリをインストールすると，GLMM のパラメータ (この場合は β と σ) を数値的な最尤推定が可能になります．R を起動してデータを読みこみ (このデータは `d` というオブジェクトだとします),

```
> library(glmmML)
> glmmML(cbind(y, 10 - y) ~ 1,
+ data = d,
+ family = binomial,
+ cluster = d$plant.ID)
```

と命じるだけで，いろいろな推定値，全個体共通部分の $\hat{\beta} = -0.0358$ (個体差ゼロの個体の結実確率は $q(\hat{\beta}) \approx 0.491$)，そして個体差のばらつき $\hat{\sigma} = 1.37$ などが得られます (なお，この例題における “真の値” は $\beta = 0,\ \sigma = 1.5$)．これらを使って，観察データに統計モデルの予測

$$100 \int_{-\infty}^{\infty} f(y \mid q(\hat{\beta} + \alpha)) g_\alpha(\alpha \mid \hat{\sigma})\ d\alpha,$$

を重ねてみると図 6 のようになります．

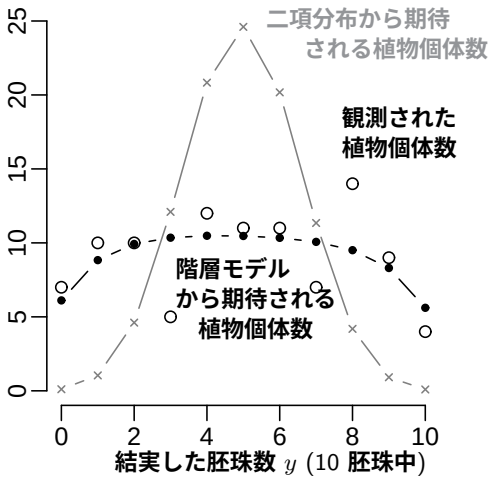

図 6 観察結果と「個性」考慮モデル

結実確率に個体差を加えることで，観察されたパターンをよりよく説明できているような予測が得られました．つまり統計モデルの改善によって，「結実確率 $q(z)$ は 0.5 ぐらい」という全体の平均だけでなく，「$z = \beta + \alpha_i$ とすると，結実確率の個体差 α_i 全体のばらつき σ は 1.4 ぐらい」という個体差に関する知識も新しく獲得できました．必要とあれば，観察データと推定値 $\{\hat{\beta},\ \hat{\sigma}\}$ を組みあわせることで図 5 のような個体ごとの α_i の事後分布も計算できます．

もうひとつの計算方法，MCMC について簡単に紹介します．個体差を考慮したモデルがこの例題のように常に簡単なものであればよいのですが，現実のデータ解析ではもっとあちこちに個体差などが入った統計モデルを構築しなければならない

状況が多くなります．個体差 α_i のたぐいがたくさん入った統計モデルでは α_i たちの積分計算は (計算量が増大するので) 事実上不可能になり，経験ベイズ法ではパラメータの推定計算ができません．

より複雑な階層ベイズモデルをあつかう場合には MCMC 法によって，パラメータの事後分布をサンプルしていく方法が使われます [4][5]．この解説では詳細は説明しませんが，`http://hosho.ees.hokudai.ac.jp/~kubo/ce/SuSemi2007.html` から推定計算のプログラム例をダウンロードできるようにしています．

個性の生態学と統計学

最後に少しだけ生態学的な話をしてみます．この解説では，図 2 に示されている架空植物の種子数の観察データにみられるパターンを説明するために個体 i の個体差 α_i を考慮した統計モデルを開発してきました．この個体差の正体って何なのでしょうか?

もしこれが架空植物ではなく，何か現実の植物だとすると，観察した個体ごとに体の大きさや年令が違っているとか，個体ごとにもっている遺伝子がちがっているという可能性はあります．これはいかにも個体差らしい要因です．ところが他にも要因はいろいろと考えることができて，たとえばその植物個体が育っている場所の明るい・暗い，あるいは土壌中の栄養塩類の多い・少ないで結実確率がちがっているのかもしれません．そうだとすると，これは個体差というよりむしろ場所差みたいなものでしょう．ともあれこういった個体差の原因になりそうなあれこれをことごとく観察データ化するなんてことは不可能です．

とくにこのような野外科学では観察できる項目が限定され，その観察の方針は「どういうパターンを説明したいのか? そのためには，(これまでの知見から) どの要因が重要そうであり観察すべきなのか?」といった考えにもとづいています．だからといって，測定できなかった他のすべてを「なんでも平均しとけばいいんでしょ?」なる発想で “なかったこと” にしてしまうのはまずいでしょう，というのがこの解説で説明したかったことです．

階層ベイズモデルの発展によって，“なかったこと” にされていた個体の差や場所の差が巧妙に統計モデル化できるようになってきました．個体差を考慮した統計モデルの利点のひとつは，(マクロな) 説明要因の推定と意味がより明確になることです (今回の例では平均的な結実確率はどう計算しても 0.5 ぐらいでしたが，個体差無視がパラメータの推定を偏らせる場合もあります [6])．それだけでなく，いままで “なかったこと” あつかいだった (ミクロな) 個体の差・場所の差などの事後分布も推定できてしまいます．こういった個性・個体差・場所差の推定結果の中にも，何か新しい発見につながる出発点があるのかもしれません．

参考文献

[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, 2004

[2] Crawley, M.J., *Statistics: an introduction using R*. John Wiley & Sons, 2005

[3] R Development Core Team, `http://www.R-project.org/`, R は誰でも無料で自由に使えるフリーな統計ソフトウェアです

[4] 伊庭幸人, 『ベイズ統計と統計物理』, 岩波書店, 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, 2005

[6] 久保拓弥・粕谷英一, 2006, 「個体差」の統計モデリング. 『日本生態学会誌』 56: 181-190 (`http://eprints.lib.hokudai.ac.jp/dspace/handle/2115/26401` からダウンロード可能)