Bayesian Inference: With ecological applications

(W.A. Link and R.J. Barker. 2010. Academic Press)

第 8章 Hidden data models

この章最後 (pp.186–200) の担当: 久保拓弥　kubo@ees.hokudai.ac.jp

8.5 Finite population sampling (有限母集団のサンプリング)

paragraph #1　母集団 (population) の一部からデータをあつめる標本調査 (sample survey) が基本的なデータ取得方法

母集団の全体像を推定するために．たとえば，平均 μ で標準偏差 σ の無限集団から m 回のランダムサンプルの平均 $\bar{x}$ のばらつきの推定量 (estimator) は $\sigma/\sqrt{m}$ となるが，サイズ M の有限集団の場合は

$$\mathrm{SE}(\bar{x}) = \frac{\sigma}{\sqrt{m}} \times \sqrt{\frac{M-m}{M-1}}$$

となり，新しく追加された部分は有限集団の補正である /* (久保) m が M に近づくにつれ $\mathrm{SE}_{\bar{x}}$ がゼロにちかづく— 標本平均 $\bar{x}$ が真の平均となる */ ．

paragraph #2　わかりやすいモデリング — 有限集団サンプリングにおけるベイズ統計モデルの利点

というのも，有限集団からサンプリングした割合を考慮した補正が自動的に入るから．とくに階層的なサンプリング設計 (hierarchical sampling design) では有効である，というのも頻度主義的にこういう状況をあつかうのは困難だから．

paragraph #3　あたかも欠測データ (missing data) モデルのようにあつかう in Rubin (1976)

詳しくは Geleman *et al.* (2004) を見よ．以下に具体例を説明する．

8.5.1 Muskrats: A simple sample survey

paragraph #1　Williams *et al.* (2002) のマスクラット *Ondatra zibethius* の個体数調査

湿地で 2 ha 調査区 50 個をつくり，そのうち 10 個をランダムに選んでマスクラットの巣 (muskrat house) の数を調べた (**Table 8.4**).

paragraph #2　頻度主義的なマスクラットの巣数推定

全 100 ha の巣数 N を知りたい．M を全調査区数 (50)，m を調べた調査区数とすると，

$$\hat{N} = M\bar{y} = 50 \times 12.1 = 605$$

となり，その標準誤差は

$$\mathrm{SE}(\hat{N}) = \sqrt{\mathrm{Var}(\hat{N})} = \sqrt{M^2 \frac{s_y^2}{m}\left(1 - \frac{m}{M}\right)} = 48.28$$

となる /* (久保) で，このあとベイズな統計モデリングのハナシとなり，上述のようなめんどうな推定量の導出にアタマを使わなくてもよい，といった展開となる */ ．

8.5.1.1 CDL, DAG, and Bayesian Model for Muskrat Data

paragraph #1 変数の定義: $\boldsymbol{y}$ と $\boldsymbol{I}$

有限集団のモデルを記述する変数: $\boldsymbol{y} = \{y_1, y_2, \cdots, y_M\}$ は各調査区内の巣の個数 /* (久保) 調べていない場所も含む */ ，$\boldsymbol{I} = \{I_1, I_2, \cdots, I_M\}$ は調査した (巣の個数を数えた) かどうかの変数 (inclusion variable) で $I_j = \{0, 1\}$ それぞれ調査しなかった・したをあらわす．したがって $\boldsymbol{y}^{\mathrm{obs}} = \{y_j | I_j = 1\}$ で $\boldsymbol{y}^{\mathrm{mis}} = \{y_j | I_j = 0\}$ となる．ここでは $f(y)$ という総数を知りたいわけで，マスクラット調査の場合は $f(y) = \sum_{j=1}^{M} \equiv N$ である．

paragraph #2 $\boldsymbol{y}$ と $\boldsymbol{I}$ の同時分布 (joint distribution)

……が有限集団の統計モデリングに必要である．**Fig. 8.7** がこの統計モデルの DAG で，parameter vector $\boldsymbol{\theta}$ は $\boldsymbol{y}$ が標本となるような確率分布を制御し，parameter vector $\boldsymbol{\phi}$ は $\boldsymbol{I}$ を (そして結果的には $\boldsymbol{y}^{\mathrm{obs}}$ も) 決めるようなサンプリングを制御している．

paragraph #3 完全データ尤度

CDL (complete data likelihood) は

$$\mathcal{L}(\boldsymbol{y}^{\mathrm{mis}}, \boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{I}) \propto [I, y | \boldsymbol{\theta}, \boldsymbol{\phi}] = [I | \boldsymbol{y}, \boldsymbol{\phi}][\boldsymbol{y} | \boldsymbol{\theta}]$$

となる /* (久保) 一番右は分割してるだけ */ ．単純なサンプリングでは $[I|\boldsymbol{y}, \boldsymbol{\phi}][y|\boldsymbol{\theta}]$ は $\boldsymbol{y}$ には依存しないだけでなく，I は確率変数ですらない /* (久保) あらかじめ 50 個のうち 10 個調べるときめているし，調査区の配置の影響とかも考慮してないため */ ．したがって CDL は $[\boldsymbol{y}|\boldsymbol{\theta}]$ に比例している．

paragraph #4 $[\boldsymbol{y}|\boldsymbol{\theta}]$ の指定

巣の個数 $\boldsymbol{y}$ はカウントデータなので，ポアソン分布か負の二項分布が妥当だろう．平均と分散が同じぐらいなので /* (久保) $\bar{y} = 12.1$, $S_y^2 = 11.6$ */ ，ポアソン分布とする．

Williams *et al.* (2002) では確率分布を指定してないので，確率にもとづく推定ではなく，漸近的な頻度計算 (asymptotic frequently calculations) によっている．これについてはあとで議論する．

8.5.1.2 Bayesian Analysis of the Muskrat Data

paragraph #1　このモデルの BUGS code (Panel 8.8)

ポアソン分布の平均 μ はガンマ分布のあいまい事前分布 (vague prior　/* (久保) あるいは無情報事前分布 */) を指定している.

paragraph #2　得られた結果

Williams *et al.* (2002) とだいたい同じになった. **Panel 8.8** をみればわかるように Bayesian のほうが容易. 推定したい全 50 ha 中の合計巣数 N の推定ばらつきは, 観測しなかった区画の個数に依存している.　/* (久保)　ここでやっていることは，パラメーター mu を推定しつつ，同時に観測されなかった 40 調査区に巣を生成するシミュレイション */

8.5.1.3 Finite- versus Infinite-Population Inference

paragraph #1　個々の区画の集団サイズの平均に興味がある場合も

それを $\bar{Y} = \frac{1}{M}\sum_{i=1}^{M} y_i = \frac{N}{M}$ と書いてみよう. 事後サンプルどうしの割算で $\bar{Y}$ の事後分布が得られる (**Fig. 8.8**).

paragraph #2　$\bar{Y}$ と μ の区別が重要

μ はマスクラットの巣の数についての無限集団みたいなものを想定しているが, $\bar{Y}$ は 50 個の平均値である. データが 10 個の場合は μ よりも $\bar{Y}$ についてよい推定が得られる. $\bar{Y}$ のばらつきのほうが 10% ほど小さく, これは最初のほうにでてきた $\sqrt{1 - m/M} = \sqrt{0.8}$ を反映している　/* (久保) $\sqrt{0.8}$ は 0.9 ぐらい */ .

paragraph #3　8.3 節であつかった占有モデルの有限・無限集団

でも同じことが生じている.

8.5.1.4 Asymptotic Normality (漸近正規性)

paragraph #1　Panel 8.8 のコードはもっと簡単になる

二番目の for ループ /* (久保) 観測しなかった区画でのポアソン乱数発生 */ は以下のように短縮できる.

```
Mu.mis.total <- (M - m) * mu
```

```
y.mis.total <- dpois(Mu.mis.total)
```

/* (久保) ポアソン乱数の和は上のようなポアソン分布にしたがうから */

paragraph #2 中心極限定理 (central limit theorem) が利用できるなら

観察していない区画の合計個体数は，平均 $(M-m)\mu$ で分散 $(M-m)\sigma^2$ の正規分布にしたがう /* (久保) $(M-m)$ 倍すると分散は $(M-m)^2$ 倍になりそうだが，中心極限定理よりこれを $(M-m)$ で割るので，分散は $(M-m)\sigma^2$ となる */ .

paragraph #3 もし母集団が正規分布なら……

y が正規分布 $N(\mu, \sigma^2)$ にしたがうとき，$\bar{y}$ の分布が正規分布であると考えるのは近似ではない (exact である)．この事例にあてはめると，区画内平均の $\bar{y}$ とそのばらつき s^2 も簡単化できて，それぞれ $\bar{y} \sim N(\mu, \sigma^2/m)$，$(m-1)s^2/\sigma^2 \sim \chi^2_{m-1}$ となる (**Panel 8.9**)． /* (久保) 脚注 13 */ Appendix B.9 によれば，これを χ^2 分布ではなくガンマ分布と使って表現することができて，BUGS code 中では s^2 の分布として，

$$s^2 \sim \mathrm{Ga}\left(\frac{m-1}{2}, \frac{m-1}{2\sigma^2}\right)$$

/* (久保) これは平均 σ^2 で分散 $2\sigma^4/(m-1)$ のガンマ分布 */ としている．

paragraph #4 得られた結果，このような方法がうまくいく場合，いかない場合

漸近正規近似の **Panel 8.9** の BUGS code を実行すると，Williams *et al.* (2002) と同じような結果が得られた．この論文の頻度主義的なやりかたはおすすめではないが，漸近正規性にもとづく近似がどのようなものであるのかを知るための参考にはなる．この近似がうまくいくのは，m や $M-m$ が大きい場合，あるいは分布が正規分布の場合である．サンプル数が少ないときには使うべきではない．

8.5.2 Stratification (層別化)

paragraph #1 単純なランダムサンプリング以外の方法: 層別化とクラスタリング

これらの方法もベイズ統計モデルによって，わかりやすく表現できる．まず層別サンプリング (stratified sampling) について考えてみよう．これは集団をいくつかの部分集団 (subpopulations, strata) にわけ，それらすべてを調査する (ただし，層ごとに調査努力がことなる)．推定は層ごとと全体に対してなされ，層の水準ごとに異なるパラメーターと全体に共通するパラメーターにわけられる．このような方法は，全体を均質な部分集団 (層) に分割できる場合に使われる．

paragraph #2 Siniff and Skoog (1964): 空中写真によるアラスカのカリブー集団サイズの推定

Panel 8.10 の上の表にあるとおり．

/* (久保) あまりちゃんと説明されてないので，以下は久保の想像: */ 地区 {A, B, $\cdots$, F} の 6 区があり，

それぞれに大きさ M が異なる．M の 1/3 ぐらいがカウントする調査区数 m である．カウントしたカリブーの層ごとの平均が $\bar{y}$，その分散が S^2.

Panel 8.10 の BUGS code は **Panel 8.9** を層ごとに計算しているだけ.

paragraph #3　結果の比較

ここでも頻度主義的な方法とベイズな方法で得られた結果がだいたい同じになった.

paragraph #4　Panel 8.10 の統計モデルでよいのか

ここでは

$$[\mu_i, \tau_i] = [\mu_i|\tau_i][\tau_i]$$

としていて，$[\tau_i]$ は $\mathrm{Ga}(a, b)$ で $[\mu_i|\tau_i]$ は $N(\psi, \kappa\tau)$ といった共役事前分布である /* (久保) ここでは $\kappa\tau$ は分散の逆数 */ ．これらを無情報な事前分布とするために，$\psi = 0$ として a, b, κ を小さくすると，(κ を小さくする影響で) パラメーター μ_i の prior precision に影響が大きい /* (久保) prior precision がよくわからない */ ．**Panel 8.10** に示しているように C や E のように S^2 の大きな層では， /* (久保) κ を? */ 10^{-6} ぐらいにしても，十分に無情報とはいえない /* (久保) うーむ? */ ．

paragraph #5　標本数が大きいので漸近近似 (asymptotic approximations) も可能

で，また頻度主義的・ベイズ的方法の結果はだいたい一致．しかしながら，層 B と F では事後分布の 95% 区間に 0 を含んでしまうので (脚注 14 参照)，この近似は良くないかも．このような場合には，漸近正規近似ではなく，マスクラットの例題で最初にやったような，モデルにもとづく方法 /* (久保) シミュレイション的なやつ */ のほうが良いだろう (**Panel 8.8**).

paragraph #6　いつまでも古いやりかたをするな

(このカリブー調査には金がかかるといったハナシがあって) …… /* (久保) どうもこういう正規近似をやるなというハナシが書いてある……この節ではそればかり説明していたのに */

8.5.3 Cluster Sampling

paragraph #1　Cluster sampling は二段階のサンプリング

最初に部分集団 (subgroups; clusters) がサンプルされて，次にそのクラスターの中から個体データが集められる /* (久保) cluster sampling の訳語として集落抽出法なるものがあるけど，あまりしっくりしないのでここでは使わない */ ．ばらつきはクラスターの選びかたと，個体の選びかたによって生じる．このようにサンプリングが階層的なので，モデルもまた階層的になる.

paragraph #2　データ: ミズナギドリ (shearwater) の繁殖つがい数の推定

15 ha の小さな島の中で，10 m 四方の調査区 50 箇所をランダムに選んで，今度はその中のとりミズナギドリの巣穴の個数をカウントした．その巣穴が使われてるかも調べた．これによって島全体で使用されている巣穴の個数 (つまりこれを繁殖つがいの数と考えて) 推定することを目的と

している．

paragraph #3 データの構造とモデリング

方形区数は $m = 50$ でこれがクラスターとなり，全体では $M = 1500$ となる．応答変数は巣穴の個数で，方形区 i の巣穴数を y_i とする．ある区画で利用されている巣穴の個数 x_i は，巣穴を利用している確率 (占有確率) p_i の二項分布にしたがうと仮定する．

paragraph #4 二種類のモデルを考えよう

ひとつは占有確率 p が全方形区で同じと仮定 (**Panel 8.11** の BUGS code) `/*` (**久保**) **次のモデルでは** p_i **が** i **によってちがうと仮定** `*/` ．このモデルも，観察された区画・そうでない区画にわけて計算している．

paragraph #5 得られた結果とモデルの妥当性のチェック

島の中のつがい数は中央値 3704 で 95% 区間は [3109, 4371] と予測された．モデルの妥当性を調べるために，5.6 節で概要を説明した二段階の事後予測調査をしてみる．

最初に，どの方形区でも巣穴の占有確率が同じかどうかを調べる．検定統計量として，

$$T^{\mathrm{obs}} = \sum_{i=1}^{50} \frac{(x_i - y_i p)^2}{y_i p}$$

を使う．マルコフ連鎖の各ステップにおいて，上の T^{obs} と同時に

$$T^{\mathrm{rep}} = \sum_{i=1}^{50} \frac{(x_i^{\mathrm{rep}} - y_i p)^2}{y_i p}$$

も評価する．ただし $x_i^{\mathrm{rep}} \sim B(y_i, p)$ である `/*` (**久保**) **二項乱数のシミュレイション** `*/` ．

ポアソン分布の仮定が妥当かどうかを調べるために，

$$T^{\mathrm{obs}} = \sum_{i=1}^{50} \frac{(y_i - \lambda)^2}{\lambda}$$

と

$$T^{\mathrm{rep}} = \sum_{i=1}^{50} \frac{(y_i^{\mathrm{rep}} - \lambda)^2}{\lambda}$$

も比較する．こちらでは $y_i^{\mathrm{rep}} \sim Po(\lambda)$ である．

paragraph #6 結果の図示

Fig. 8.9 に示しているように，二項分布の仮定はダメそうで (左)，ポアソン分布の仮定はまあ妥当 (右).

paragraph #7 二項分布の仮定がうまくいかなかった理由

場所によって占有確率が異なるので，方形区に由来する random effects を考慮する必要がある．

paragraph #8　次のモデル

方形区ごとの random effects を考慮して，占有確率が p_i が i ごとに異なるようにする．

```
logit(p[i]) <- logitp[i]
logitp[i] ~ dnorm(mu.p, tau.p)

mu.p ~ dnorm(0, 1.0E-6)
sd.p ~ dunif(0, 100)
tau.p <- 1 / pow(sd.p, 2)
```

/* (久保) ようするに GLMM ベイズ版である階層ベイズモデル化 */

paragraph #9　得られた結果

つがい数の中央値が 200 減少したり予測の幅が 50% ひろがるなど，結果がいろいろ変わったけれど，**Fig. 8.10** に示しているように占有確率のモデリングはマシになった．

8.5.4 Auxiliary Variables (補助変数)

paragraph #1　区画ごとの情報をモデルに利用できる

ミズナギドリの例では方形区ごとの random effects をモデルにくみこんで現実に対応し，観察されてない方形区の予測の幅もひろがってしまった．方形区ごとの情報を共変量 (covariate) として使えばモデルがマシになるかもしれない /* (久保) つまり環境変数その他 fixed effects 的なものをくみこむ */ ．サンプリングをした方形区だけでそういう共変量が得られる場合と，その他でも得られている場合のふたつの状況が考えられる．

paragraph #2　区画が exchangeable であれば

サンプリングをした方形区でしか共変量が得られていなくても，方形区が exchangeable であると考えてよければ，その他の方形区の共変量を欠測値としてモデリングすればよい．

paragraph #3　マスクラットの巣の個数の問題にもどると……

Williams *et al.* (2002) の第二の調査では調査区の面積が異なっている．これを共変量 (面積 a_i) として，ポアソン分布の平均を $\mu_i = \lambda a_i$ としてみよう　/* (久保) これは GLM 的にいうと “offset 項わざ” */ ．

paragraph #4　得られた結果

Panel 8.12 に示したコードであてはめてみると，巣の個数の合計 N の分布は **Fig. 8.11** (上) のようになる．**Fig. 8.11** (下) T 統計量も妥当そうにみえる．Williams *et al.* (2002) は割算値の漸近正規性を仮定した区間推定をしているが，彼らの結果は N の信頼区間がせますぎるように思

える (ベイズの結果と比較すると).

paragraph #5 全区画で共変量があれば望ましいのだが

全部は無理でも，サンプリングをしない区画でも共変量を測定したほうがよいだろう…… /* (**久保**) (**略**) */

8.6 Afterword

paragraph #1 「こんなデータがあればいいのに」モデリング

第一章ではベイズは「便利だから使いたくなる」もので，その便利さとは「多くの観測データを単純にあつかえる」ことにあるとほのめかしていた．この章はまさにそういうことを示すものである．それぞれの事例で，モデリングは「こんなデータがあればいいのに」(**Table 8.5**) と考えるところから考えはじめている．見えないデータ (latent data) は無限集団からのサンプリングと考えると，無限集団のパラメーターについても，有限の標本の統計量についても推測が可能になる．

paragraph #2 ベイジアンな方法とはデータの再構築

この章のそれぞれの事例の例題のデータは，条件つき分布 [観測データ — 潜在データ] の観点からすると,「こんなデータがあればいいのに」版データの圧縮もしくは不完全版であった．頻度主義的なデータ解析は ODL (observed data likelihood) を使ってなされていて，これは潜在データの事前分布を積分することで得られたものである．ODL はめんどうになりがちで，観測データの同時分布にもとづく CDL のほうがマシ．ベイズモデリングはこれにもとづいている．ベイズの maramatanga のもとでは，未知の数量だろうがパラメーターだろうが欠測値だろうが，すべてひとしくあつかわれる．この章では，事後分布からのサンプリングによって，再構築された「こんなデータがあったらいいな」を解析した．