最近のベイズ推定研究の小地域の人口動態指標推定への応用の研究

中田　正（ナカダ タダシ）＊1　齋藤　重正（サイトウ シゲマサ）＊2　六車　史（ムグルマ フミト）＊3

目的　人口動態統計のような全数調査であっても存在すると考えられる「モデル誤差」の概念を導入し，小地域の指標推定において，最近のベイズ統計学の手法を用いることで，モデル誤差を克服することを目的とした。

方法　平成10～14年の高知県における市区町村別標準化死亡比について，二次医療圏ごとにベイズ推定した結果および県全体でベイズ推定した経験ベイズ推定値を算出し，ベイズ推定しない結果と比較した。

結果　ベイズ推定することで，小地域間の偶然変動によるばらつきを，かなり小さくできるという結果を得たが，地域選定の判断基準を得るまでは至らなかった。

結論　コンピューターの処理能力の向上により，厳密な計算はできなくともシミュレーションにより近似値を得る方法（MCMC 法）が開発され，ベイズ推定の応用範囲が事前分布の制約を受けないところまで拡大したが，シミュレーション結果の妥当性には注意が必要である。小地域人口動態指標における地域選定の考え方として，より広い地域を設定すれば得られる統計指標は安定するが，小地域特性の反映度は薄くなるので，行政的に意味のある結果が得られるよう設定することが重要である。今後の課題として，地域選定の判断基準の分析と，新たな指標を設定し，ベイズ推定法等による計算とその妥当性について，さらに研究する必要がある。

キーワード　小地域，人口動態指標，モデル誤差，市区町村別標準化死亡比，ベイズ推定

Ⅰ　は　じ　め　に

厚生労働省大臣官房統計情報部が行っている統計調査には，大きく分けて，

①　標本調査

②　全数調査

の２種類がある。さらに，全数調査においても全国，都道府県別のように比較的広域で集計した統計と，市区町村別のように小地域で集計した統計がある。これら統計調査の調査結果を，「実態を表す」ものとして公表している基本的な考え方は次のとおりである。

（１）　標本調査

標本調査の目的は，母集団（全数）に関して知りたい事柄があるとき，その一部である標本に対してその事柄を調べることで，「母集団の実態」を推定することである。今，母集団に関して知りたい事柄が，ある属性値 x であったとすると，標本調査で知りたい「真の値」とは，たとえば x の母平均 μ である。これら母集団に関する未知の指標を，確率変数である標本平均 $\overline{X}$ や標本分散 s^2 の実現値を頼りに推定するのである。

調査の対象となる標本は無作為抽出に基づき

＊１日興フィナンシャル・インテリジェンス副理事長

＊２厚生労働省大臣官房統計情報部人口動態・保健統計課長補佐　＊３同企画課審査解析室主査

得られたものであるが，生じ得るすべての標本からなる標本空間Ωを考えると，それは空間Ωの一点ω∈Ωが実現したものとみなせる。例えば標本平均 $\overline{X}$ は，Ωを構成する個々の標本ωごとに値が一意に定まる関数

$\overline{X}(\omega)：\Omega\rightarrow\mathbb{R}$

として，Ω上の確率変数となっている。そして，調査結果から得られる標本平均に対して，確率変数 $\overline{X}$ が従うと考えられる確率分布を仮定し，その確率分布の平均や分散を推定するのである。すなわち，標本調査の結果として提示される「実態」とは，母集団全体ではなくその一部である標本を調べた結果である以上，常に「標本誤差評価付きの母数の推定値」であるということになる。

具体的には，標本の大きさを n とし，標本を構成する各個体の属性値を X_1，X_2，…，X_n とすると，無作為抽出である以上，X_1，X_2，…，X_n は互いに独立に同一の分布（属性値 X の母集団分布）に従う（independently and identically distributed; i.i.d.）確率変数であると仮定される。このとき標本平均 $\overline{X}$ は，標本ω＝（X_1，X_2，…，X_n）の実数値関数として

$$\overline{X}(\omega)=\frac{1}{n}\sum_{i=1}^{n}X_i$$

で定義され，$\overline{X}$ 自身もまたひとつの確率変数となる。また，

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

で定義される標本分散 s^2 も同様にひとつの確率変数となる。

標本平均 $\overline{X}$ および標本分散 s^2 は，母集団分布によらずその期待値が，

$E(\overline{X})=\mu$，$E(s^2)=\sigma^2$

と，母平均μおよび母分散 σ^2 にちょうど等しくなる（不偏性）意味において，母集団と標本とをつなぐ重要な量である。また，母集団の大きさを N と置くと，標本平均 $\overline{X}$ の分散は，

$$V(\overline{X})=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\cong\frac{\sigma^2}{n}$$

（$\frac{n}{N}\ll 1$ のとき）

と計算されるので，$n\rightarrow\infty$ のとき $V(\overline{X})$ は０に近付いていくことがわかる。このことは，標本の大きさ n が十分大きいときには，確率変数 $\overline{X}$ の実現値が母平均μの周りに集中し，そこから大きく外れることがほとんどなくなるであろうことを意味する。すなわち，複数の個体からなる標本の平均値を取ることにより，母平均を推定する精度を上げているのである。

標本調査の実務上問題となるのは，

ア　無作為抽出であること（標本設計および標本抽出）

イ　調査票が確実に回収されること（回収率）

ウ　回収された調査票に，記入ミス等の事務処理上のミスがないこと

であるが，アに起因する誤差を標本誤差，イまたはウに起因する誤差等標本誤差以外の調査誤差を非標本誤差という。公表されているのは，アの標本誤差である。

（２） 全数調査

一方，全数調査の場合には，全数を調べる以上，一部しか調べない標本調査と違って標本誤差は存在せず，非標本誤差を最小にするだけでよいと考えられており，したがって調査結果がそのまま「実態を表す」ものとされる。

たしかに，全国，都道府県単位のように比較的広域の統計では，実態を表すと考えてよいであろう。しかしながら，市区町村単位ではどうであろうか。特に発生率などの指標を計算する場合はどうであろうか。

（３） 全数調査におけるモデル誤差の存在

現在の出生率（人口対比）は，およそ0.01（＝１％）である。したがって，規模の小さい市区町村，例えば1,000人程度の規模だと，単純計算では年に

1,000人×0.01＝10人

程度の出生が期待されるところである。このことは観点を変えると，人口1,000人の市区町村において年間10人の出生があったならば，その市区町村の出生率は0.01と計算されることを意味する。ここで統計数値としてすでに得られて

いるのは，人口の1,000人と出生数の10人であって，出生率0.01は統計数値から計算される指標ということになる。しかしながら人口動態統計の結果をみると，例えば，平成16年には出生数０という町村が存在しており，実績はそのとおりであっても，これらの町村の「実態」を表す指標として，そのまま出生率が０であるとしていいかどうかは，疑問のあるところである。出生数を死亡数に置き換えてみると，その意味するところがよりはっきりするであろう。出生や死亡を，偶然変動に左右される確率事象ととらえると，その真の確率は，少ない結果から計算される頻度割合とは大きくずれることがあると考えるのは合理的である。

このように母集団が小さい場合，標本誤差や非標本誤差とは別の誤差を考える必要があることがわかる。特に，特定の死因による死亡率のように，出現率がきわめて小さい事象の指標を計算する場合には，誤誘導する恐れがあるからである。調査に起因する標本誤差，非標本誤差のような「調査誤差」に対して，この誤差を仮に「モデル誤差」ということにすると，全数調査の場合，母集団が小さいとき，すなわち小地域のときモデル誤差をどう考えるかが重要となってくる。

- 誤差
  - 調査誤差
    - 標本誤差
      - ・抽出調査により全数を調べないことによる誤差
    - 非標本誤差
      - ・回収率等による誤差
      - ・全数調査でも抽出調査でも生じうる
  - モデル誤差
    - ・全数調査であっても，発生率の調査には，本来の発生率と考えられる水準とは異なる実績が得られるという誤差がある。特に小地域，小発生率の場合著しい。

（４） ベイズ統計学によるモデル誤差の克服

厚生労働省統計情報部では，「平成10～14年人口動態保健所・市区町村別統計」において，

○市区町村別合計特殊出生率

○市区町村別（全死因）標準化死亡比

について，前記のモデル誤差を克服するため，２つの工夫を行っている。

１つは，これらの市区町村別指標を算出するのに，単年度ではなく，国勢調査年（平成12年）を中心とする５年分のデータを用いることである。これは，同一の分布（平均μ，分散σ^2）に従う互いに独立な確率変数（１年間の出生数または死亡数）$X^{(H10)}$，$X^{(H11)}$，$X^{(H12)}$，$X^{(H13)}$，$X^{(H14)}$について，その５年分の平均

$$X=\frac{1}{5}(X^{(H10)}+X^{(H11)}+X^{(H12)}+X^{(H13)}+X^{(H14)})$$

を取ることで，

$$E(X)=\mu,\quad V(X)=\frac{\sigma^2}{5}<\sigma^2$$

となり，単年度よりも分散が小さくなり，発生件数が安定することから，小地域指標に特有のモデル誤差を小さくする効果があると期待される。なお，国勢調査年を中心として，長期間のデータをとればとるほどいいかというと，そのような訳にはいかない。それは，長期間観察すると，保健衛生の状況や意識の変化といった構造変化が起きている可能性があり，もはや同一の分布に従うとは考えられず，平均をとったとき構造変化分を織り込んでしまう恐れがあるからである。

さらに，もう１つの工夫として，当該市区町村を含む二次医療圏地域で得られた出生率，死亡率を事前分布（のパラメータ）とし，当該市区町村の出生率，死亡率を尤度として事後分布（のパラメータ）を求めるベイズ推定により算出している。これは，ある市区町村の発生率は，おおむねその市区町村を含む二次医療圏地域での発生率に近いという前提を置き，その市区町村で得られた実績に基づき，二次医療圏地域平均の発生率をいわば補整して，その市区町村の本来の発生率にするというものである。

なお，平成12年市区町村別生命表においても同様なベイズ推定を行っている。

（５） 新たな指標の必要性と生じる困難

「平成18年医療制度改革」によれば，都道府県医療費適正化計画において，生活習慣病対策は重要な柱の１つと位置づけられている。また，都道府県内の国保保険者（市区町村）等の保険者には健診や保健指導の義務が課されることとなったため，生活習慣病による有病，罹患および死亡の実態把握は，計画作成，実施，点検・評価および見直し・改善の一連の循環（PDCAサイクル）において重要な位置を占める。このため，市区町村別の標準化死亡比（SMR）について，全死因のみならず，生活習慣病を中心とする主要死因別SMRの公表が求められるようになった。

この場合に問題となるのは，

① 最大の割合を持つがん（悪性新生物）でも全死因の約３割と少ないこと

② 既に公表している全死因に関するSMR（ベイズ推定値）と，死因ごとに算出したSMR（ベイズ推定値）について，加法性が成り立つとは限らないこと

である。

②にいう加法性とは，ベイズ推定した死因別SMRを合計すれば全体のSMRになることをいう。しかしながら，SMRについては，各市区町村の全死因のSMRがベイズ推定されているが，各死因別SMRをベイズ推定し，分子分母をそれぞれ合計しても全死因のSMRのベイズ推定値に一致するとは限らない。このような状況下でも，ベイズ推定の手法が妥当かどうか，改善する必要はないかどうか，検討する必要がある。

（６） 研究の目的

しかしながら，近年の研究やコンピューターの進歩に対応し，以前では評価が困難であった対象に関しても適用可能な，MCMC法（マルコフ連鎖モンテカルロ法）などの数値計算法が急速に発展してきている。このため，これまで小地域の人口動態指標における研究において解決が困難であったこれらの課題などに対しても，最新の手法による解決の可能性がでてきている状況にある。本研究は，まずベイズ統計学導入の動機となったモデル誤差の考察を行い，ベイズ統計学を取り巻く最新の研究事情やその成果を調査するとともに，このような新たな手法を，小地域の人口動態指標の推定に応用する方法に関して研究したものである。

Ⅱ モデル誤差の考察

（１） 有限集団におけるモデル誤差

要素数Nの集団において，各要素における発生率がp（$0<p<1$）である事象（出生や死亡）については，発生数rは，二項分布$B(N, p)$に従うと想定しうる。

$${}_NC_r p^r (1-p)^{N-r}$$

発生率$\theta = \frac{r}{N}$は，0，$\frac{1}{N}$，$\frac{2}{N}$，…，$\frac{N}{N}=1$のいずれかの値をとり，特定の$\frac{r}{N}$をとる確率は，${}_NC_r p^r (1-p)^{N-r}$である。このとき，θの期待値はp，分散は$\frac{p(1-p)}{N}$となるので，変動係数は，

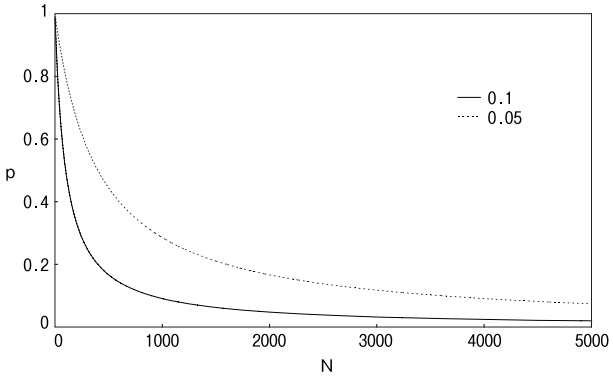
$$\frac{\sqrt{\frac{p(1-p)}{N}}}{p} = \sqrt{\frac{1-p}{Np}}$$

と計算される。これはNについても，pについても減少関数であるので，

ア 十分大きな集団である（$1 \ll N$）

イ 十分発生率の高い事象である

図1 関数f（N, P）の等高線



$(0 \ll p)$

のいずれも満たされる場合には，変動係数が小さくなり，観測値θの安定性は高いが，アかイのいずれか，あるいは両方満たさない場合には，観測値θの安定性は低くなる。このことをみるため，関数

$$f(N,\ p)=\sqrt{\frac{1-p}{Np}}\quad(0<p<1)$$

の等高線を調べる。$f(N,\ p)=0.05,\ 0.1$とすると，図１のとおりとなる。

$N=1000$程度だと，$p=0.1$でも１割程度のモデル誤差がある。また，$p=0.05$だと，$N=5000$でも１割程度のモデル誤差がある。また，観測値を$\hat{p}$とすると，チェビシェフの不等式から，一般に，

$$P\left\{|p-\hat{p}|>n\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right\}\leqq\frac{1}{n^2}$$

となる。すなわち，有意水準12%程度で，

$$\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\leqq p\leqq\hat{p}+3\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

が成り立つ。

$$\hat{p}-3\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\leqq\frac{\hat{p}}{2}$$

とすると，

$$N\leqq 36\times\frac{1-\hat{p}}{\hat{p}}$$

であるから，$\hat{p}=0.01$を代入すると，

$$N\leqq 3546$$

すなわち，$\hat{p}=0.01$の事象について，$N=3500$程度だと$p=0.005$（0.5%）と判断することもありうるということになる。次に，rについて考えると，rの期待値はNp，分散は$Np(1-p)$となり，変動係数は，

$$\frac{\sqrt{Np(1-p)}}{Np}=\sqrt{\frac{1-p}{Np}}$$

したがって，$N=1000$，$p=0.1$とすると$Np=100$であるが，rについては１割程度のモデル誤差がある。90～110程度ということになる。

（２） 無限集団におけるモデル誤差

（１）でみたように，Nが大きいとき，変動係数が大きくなるのはpが小さいときである。そこで，$\lambda=Np$を一定とし，$N\to\infty$，$p\to 0$とした極限を考えると，二項分布$B(N,\ p)$はポアソン分布$Po(\lambda)$に近づく。事象がn個発生する確率は，

$$\frac{\lambda^n}{n!}e^{-\lambda}$$

であり，$Po(\lambda)$に従う。事象の発生数の期待値はλ，分散はλであるので，変動係数は，

$$\frac{\sqrt{\lambda}}{\lambda}=\frac{1}{\sqrt{\lambda}}$$

となる。$N=1000$，$p=0.1$としたとき，

$$\lambda=Np=100$$

よって変動係数は，

$$\frac{1}{\sqrt{100}}=0.1$$

となる。したがって，$N=1000$の場合，事象の発生数には常に１割のモデル誤差があることになる。

人口動態統計では全国の年齢階級別統計や都道府県別統計を月報や年報で公表しているが，これは$N\to\infty$に当たる。したがって，pが極めて小さい事象についてはポアソン分布に従って事象が発生していると考えることができるため，全数統計であるにもかかわらず，発生件数は年によって変動することになる。

（例）$N=300,000$，$p=0.001$の事象の発生率

二項分布$B(N,\ p)$をポアソン分布$Po(Np)$により近似すると変動係数は，

$$\frac{1}{\sqrt{300}}\fallingdotseq 0.0577$$

となる。したがって，６％弱のモデル誤差があることになる。計算の都合上，二項分布$B(N,\ p)$を正規分布$N(Np,\ Np(1-p))$により近似することとすると，上記の事象の発生件数は，$N(300,\ 299.7)$に従う。したがって，有意水準５％で

$$300-2\sqrt{299.7}\leqq X\leqq 300+2\sqrt{299.7}$$

$$265.4\leqq X\leqq 334.6$$

となる。１割近くぶれることとなり，実績が得られたとき件数としては，百の位はあてにできても，十の位は既にあやしいこととなる。発生率は

$$N\left(p, \frac{p(1-p)}{N}\right)=N(0.001, 3.33\times10^{-9}),$$

$$\sigma=\sqrt{\frac{p(1-p)}{N}}=5.77\times10^{-5}$$

有意水準５％では

$0.000885 \leqq X \leqq 0.00115$

このことは，発生率としては0.0009～0.0011程度ということになり，１桁目をとって0.001とみるのがよく，その下の２桁目以下はあまり意味がない。偶然変動による可能性が大きいからである。

Ⅲ　ベイズ統計学の概況

（１）　ベイズの定理とベイズ推定

確率空間（Ω，$\boldsymbol{B}$，P）においてA，$B\in\boldsymbol{B}$とする。事象Aが生起したときに，事象Bが生起する条件付き確率を$P(B\mid A)$と書くと，

$$P(B\mid A)=\frac{P(A\cap B)}{P(A)}=\frac{P(A\mid B)P(B)}{P(A)}$$

が成り立つ。また，パラメータ空間Θにおけるパラメータθの分布の確率密度関数を$p(\theta)$，データ空間Ωに値をとる確率変数Xの分布の確率密度関数を$f(x)$とする。パラメータがθをとるときのXの分布の確率密度関数を$f(x\mid\theta)$とすると，$X=x$のときのθの分布の確率密度関数$p(\theta\mid x)$は，

$$p(\theta\mid x)=\frac{f(x\mid\theta)p(\theta)}{\int f(x\mid\theta)p(\theta)d\theta}$$

で与えられる。これらをベイズの定理という。

ベイズ統計学では上記の数学の定理を用いて，推測の論理を以下のように組み立てる。未知パラメータθの事前分布$p(\theta)$のもとでデータxが観測されていたとすると，θの事後分布$p(\theta\mid x)$が

$$p(\theta\mid x)=\frac{f(x\mid\theta)}{\int f(x\mid\theta)p(\theta)d\theta}p(\theta)$$

として得られる。

（事後分布）＝（尤度）×（事前分布）

事前分布$p(\theta)$は，推測しようとする者が決めるものであり，いったん決めると，観測データxによる補整係数（尤度）により，事後分布$p(\theta\mid x)$が得られる。

（２）　full–Bayes 法と empirical Bayes 法

$p(\theta)$の決め方には２通りある。１つは，パラメータθの事前分布$p(\theta)$に含まれる超パラメータについて情報がないという状況を表す確率分布である「無情報事前分布」を仮定するfull-Bayes 法である。無情報事前分布としては，一様分布が自然であるが，定義域が無限である場合には“ほとんど一様分布”であるような分布をとる。例えば，超パラメータαの範囲が（$-\infty$，∞）のときは，

$\alpha\sim N(0,\ \sigma^2)$　$\sigma\cong100$

［0，∞）のときは，

$$\frac{1}{\alpha}\sim G(a,\ a)\quad a\cong0.001$$

ここでGはガンマ分布を表し，$G(a,\ b)$は

$$g(x\mid a,\ b)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$$

となる確率分布である。

もう１つは，パラメータθの事前分布$p(\theta)$に含まれる超パラメータの事前分布は未知と考え，それを観測データから最尤推定するものでempirical Bayes 法といわれている。例えば

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

のとき，超パラメータμ，σ^2を，観測データから最尤法で推定するものである。

ベイズ統計学と非ベイズ統計学の違いをみるために，例で比べてみる。

（例）ある母集団から無作為に選んだ１組の標本（X_1，X_2，…，X_n）に対して，母集団分布が$N(\mu,\ \sigma^2)$であるとして，μ，σ^2を推定することを考える。

非ベイズ統計学による方法

標本平均$\overline{X}=\frac{1}{n}(X_1+X_2+\cdots+X_n)$を考えると，$E(\overline{X})=\mu$より，

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}$$

また，標本分散$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$を考えると，$E(s^2)=\sigma^2$より，

$$\widetilde{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

ベイズ統計学による方法

２つの未知パラメータμ，σ^2の事前分布$p(\mu,\ \sigma^2)$を設定し，事後分布

$p_1(\mu\mid x_1,\ x_2,\ \cdots,\ x_n)$,

$p_2(\sigma^2\mid x_1,\ x_2,\ \cdots,\ x_n)$

を推定する。以下，$\boldsymbol{x}=(x_1,\ x_2,\ \cdots,\ x_n)$と置く。パラメータとデータの同時確率分布は

$$p(\mu,\sigma^2,\boldsymbol{x})=p(\mu,\sigma^2)\prod_{i=1}^{n}f(x_i\mid\mu,\sigma^2),$$

$$f(x_i\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

となるので，$p_1(\mu\mid\boldsymbol{x})=\int p(\mu,\ \sigma^2,\ \boldsymbol{x})d\sigma^2$より，

$$p_1(\mu\mid\boldsymbol{x})\propto\int p(\mu,\sigma^2)\prod_{i=1}^{n}f(x_i\mid\mu,\sigma^2)d\sigma^2,$$

また$p_2(\sigma^2\mid\boldsymbol{x})=\int p(\mu,\ \sigma^2,\ \boldsymbol{x})d\mu$より

$$p_2(\sigma^2\mid\boldsymbol{x})\propto\int p(\mu,\sigma^2)\prod_{i=1}^{n}f(x_i\mid\mu,\sigma^2)d\mu$$

事前分布として

$$\mu\sim N(0,\ \sigma_\mu^2),\ \ \tau=\frac{1}{\sigma^2}\sim G(a,\ a)$$

を仮定する。σ_μ^2，aが超パラメータであり，無情報事前分布すなわち一様分布（＝定数）とする（full-Bayes法）。$\tau_\mu=\frac{1}{\sigma_\mu^2}$とすると，

$$p_1(\mu\mid\boldsymbol{x})\propto N\left(\frac{n\overline{x}\tau}{\tau_\mu+n\tau},\ \frac{1}{\tau_\mu+n\tau}\right),$$

$$p_2(\sigma^2\mid\boldsymbol{x})\propto G\left(a+\frac{n}{2},\ a+\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right)$$

事後分布$p_1(\mu\mid\boldsymbol{x})$，$p_2(\sigma^2\mid\boldsymbol{x})$の期待値が（μ，$\sigma^2$）のベイズ推定値であり，

$$\widetilde{\mu}=\frac{n\overline{x}\,\widetilde{\tau}}{\tau_\mu+n\widetilde{\tau}},$$

$$\widetilde{\sigma}^2=\frac{2a+\sum_{i=1}^{n}(x_i-\widetilde{\mu})^2}{2a+n}=\frac{1}{\widetilde{\tau}}$$

ここから$\widetilde{\mu}$，$\widetilde{\tau}=\frac{1}{\widetilde{\sigma}^2}$を解けばよいが，$n\to\infty$とすれば

$$\widetilde{\mu}\to\widetilde{x},\ \ \widetilde{\sigma}^2\to\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

となる。

（３） 共役法（Conjugate 法）

ベイズ推定の公式

$$p(\theta\mid x)=\frac{f(x\mid\theta)p(\theta)}{\int f(x\mid\theta)p(\theta)d\theta}$$

によれば，分母に積分があらわれる。また，empirical Bayes法の例でみられるように，θはある未知の超パラメータを用いる分布族に属する分布に従うとして計算するので，$p(\theta\mid x)$も，その超パラメータを用いた分布となるが，θの期待値は，その事後分布$p(\theta\mid x)$から計算される。ベイズ推定の計算を解析的に解くことを考えると，積分

$$\int\theta\cdot p(\theta\mid x)d\theta=\int\theta\cdot f(x\mid\theta)p(\theta)d\theta$$

が解析的に計算できることが条件となってくる。$f(x\mid\theta)$は，実用上，二項分布，ポアソン分布，正規分布，指数分布となることが多いと考えられるが，上記の積分が実行できるような$p(\theta)$としては，以下のような分布を考えると都合がいい。これら$p(\theta)$の分布を共役事前分布という。

	$f(x\mid\theta)$	$p(\theta)$
①	二項分布	ベータ分布 $B(p,\ q)$
②	ポアソン分布	ガンマ分布 $\Gamma(\alpha)$
③	正規分布	正規分布
④	指数分布	ガンマ分布

いずれも$p(\theta\mid x)=f(x\mid\theta)\ p(\theta)$が$p(\theta)$と同じ分布族に属し，解析的計算が可能である。特に，θに関して積分可能なので，事後分布$p(\theta\mid x)$の期待値として解析的にベイズ推定値を計算することができる。

（４） MCMC 法（マルコフ連鎖モンテカルロ法）

確率変数Xが密度関数$\pi(x)$をもつとき，関数$f(X)$の期待値は

$$E_\pi(f(X))=\int f(x)\pi(x)dx$$

である。ベイズ推定には，この積分値を必要と

するが，解析的には計算が困難な場合がある。このような場合，$\pi(x)$ から独立な標本

$$x^{(1)}, x^{(2)}, \cdots, x^{(N)} \sim \pi(x), x^{(i)} \perp x^{(j)} if \quad i \neq j$$

が得られれば，

$$E_{\pi}(f(X)) \cong \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)})$$

で近似できる。この近似がモンテカルロ積分である。しかし，一般には複雑な $\pi(x)$ から独立な標本を得ることも困難である。このため，独立ではないものの，あるマルコフ連鎖で発生させた標本で代用することを考える。この組み合わせをマルコフ連鎖モンテカルロ（MCMC）法という。

MCMC 法の本質は，$\pi(x)$ の分布を，あるマルコフ連鎖により発生させた乱数列でシミュレーションすることである。一次元の分布で説明すると，遷移確率 $p(\cdot|\cdot)$ を持つマルコフ連鎖から乱数を発生させるとは，$x(i+1) \sim p(x \mid x(i))$ ということであり，$x(i+1)$ は $x(i)$ には依存するが，$x(0)$，$x(1)$，…，$x(i-1)$ とは独立である。具体的には，Metropolis-Hasting アルゴリズムというアルゴリズムが確立されており，このアルゴリズムにしたがって発生させた乱数列から得られる定常分布が $\pi(x)$ となることが証明されている。*

Ⅳ　ベイズ推定の人口動態指標への応用

（１）　市区町村別合計特殊出生率

出生率の事前分布としてベータ分布を選択する。母数空間を $\Theta[0, 1]$，母数（出生率）を θ とすると，事前分布の密度関数は

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

となる。ただし，$B(\alpha, \beta)$ はベータ関数である。ベータ関数の基本的性質として，確率変数 $\tilde{x}$ がベータ分布にしたがうとき，

$$\text{平均：} E(\tilde{x}) = \frac{\alpha}{\alpha+\beta},$$

$$\text{分散：} V(\tilde{x}) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

となる。観測によって人口 N と出生数 B を得るわけであるが，偶然変動を大きく受けるのは B であり，N はそれほど変動しないと考える。ここで人口は既知とし，出生数は確率変数 $\tilde{B}$ の実現値と考える。さらに出生率 θ が真のとき，$\tilde{B}$ が二項分布 $Bin(N, \theta)$ にしたがうと仮定する。このとき，$\tilde{B}$ の確率密度関数 $f(B \mid N, \theta)$ は，

$$f(B \mid N, \theta) = {}_{N}C_{B}\theta^{B}(1-\theta)^{N-B}$$

となる。これを出生率 θ の関数とみなせば尤度関数となる。

以上より，観測により N, B を得たという条件の下で，母数 θ の事後分布の密度関数を $p(\theta \mid B, N)$ とすると，ベイズの定理より，

$$\begin{aligned} p(\theta \mid B, N) &= \frac{f(B \mid N, \theta)p(\theta)}{\int_0^1 f(B \mid N, \theta)p(\theta)d\theta} \\ &= \frac{{}_{N}C_{B}\theta^{B}(1-\theta)^{N-B}p(\theta)}{\int_0^1 {}_{N}C_{B}\theta^{B}(1-\theta)^{N-B}p(\theta)d\theta} \\ &= \frac{1}{B(\alpha', \beta')} \theta^{\alpha'-1}(1-\theta)^{\beta'-1} \end{aligned}$$

（ただし，$\alpha' = \alpha + B$, $\beta' = \beta + N - B$）

このように，事後分布も α'，β' をパラメータとするベータ分布となる。事後分布の平均値および分散は，ベータ分布の基本的性質より，

$$E(\tilde{\theta} \mid B, N) = \frac{\alpha+B}{\alpha+\beta+N},$$

$$V(\tilde{\theta} \mid B, N) = \frac{(\alpha+B)(\beta+N-B)}{(\alpha+\beta+N)^2(\alpha+\beta+N+1)}$$

となる。

（２）　市区町村別標準化死亡比

母集団の母数を θ_d（標準化死亡比）とし，事前分布としてガンマ分布 $\Gamma(\alpha, \beta)$ を選択する。このとき，データの観測による市区町村

＊コンピューターでマルコフ連鎖モンテカルロ法によるシミュレーションを実行するソフトウェアについては，たとえば The BUGS Project（http://www.mrc-bsu.cam.ac.uk/bugs/）等を参照されたい。

の死亡数 d にポアソン分布を仮定し，期待死亡数を e とすると，事後分布もガンマ分布 $\Gamma(\alpha+d,\ \beta+e)$ となる。その期待値および分散は，

$$E(\theta_d \mid d)=\frac{\alpha+d}{\beta+e},\quad V(\theta_d \mid d)=\frac{\alpha+d}{(\beta+e)^2}$$

となる。

表１　高知県　標準化死亡比（平成10～14年）

	男	ベイズ推定値	経験ベイズ推定値	女	ベイズ推定値	経験ベイズ推定値
高知市	101.7	101.7	101.7	95.2	95.2	95.2
室戸市	128.8	127.3	125.5	123.9	122.8	121.0
安芸市	112.1	113.1	111.0	112.3	112.2	110.8
南国市	101.8	101.7	101.8	105.4	104.6	105.0
土佐市	104.5	103.9	104.3	115.4	112.8	113.9
須崎市	99.8	99.7	100.1	88.9	89.5	89.8
中村市	97.6	98.1	98.1	96.0	96.5	96.2
宿毛市	93.0	94.6	94.1	91.6	93.0	92.2
土佐清水市	102.9	102.3	102.9	101.1	100.7	100.8
安芸郡						
東洋町	125.5	122.5	116.7	96.1	101.0	97.0
奈半利町	119.7	119.3	113.7	117.1	115.8	111.3
田野町	107.7	112.9	105.7	77.6	89.3	86.1
安田町	135.6	127.6	121.9	137.6	128.3	120.9
北川村	86.4	105.6	95.5	93.0	102.5	96.2
馬路村	93.0	111.1	99.6	111.8	111.9	102.7
芸西村	116.8	117.6	111.7	91.3	96.2	93.4
香美郡						
赤岡町	147.0	117.2	126.6	122.6	109.6	113.6
香我美町	94.1	97.5	96.6	82.2	87.7	86.8
土佐山田町	103.3	102.9	103.2	98.5	98.2	98.5
野市町	102.8	102.4	102.9	89.3	91.0	90.8
夜須町	106.4	103.6	105.2	104.0	100.4	102.1
香北町	81.0	89.4	86.3	76.3	82.5	81.1
吉川村	124.1	106.5	111.2	146.0	109.6	116.3
物部村	92.9	97.5	96.4	97.0	96.6	97.5
長岡郡						
本山町	80.9	90.9	87.6	72.3	82.2	80.2
大豊町	112.5	108.4	110.5	94.2	94.8	95.1
土佐郡						
鏡村	96.1	100.1	100.1	109.2	99.9	102.7
土佐山村	64.6	93.3	88.4	108.3	99.4	102.2
土佐町	103.2	102.3	103.1	74.0	81.8	80.1
大川村	65.2	96.0	92.6	89.6	95.3	96.7
本川村	120.0	104.3	107.9	122.3	101.3	105.2
吾川郡						
伊野町	100.0	100.3	100.4	89.8	90.9	90.7
池川町	91.0	96.9	95.5	93.8	95.0	95.5
春野町	109.7	107.3	108.6	105.0	103.1	104.1
吾川村	98.5	100.1	100.1	85.5	90.5	90.0
吾北村	128.9	113.6	119.5	90.5	93.1	93.2
高岡郡						
中土佐町	114.8	110.5	111.8	98.2	96.6	98.3
佐川町	89.7	92.9	91.7	89.0	90.5	90.4
越知町	95.9	98.1	97.5	86.3	89.2	88.7
窪川町	104.0	103.3	103.9	92.4	92.5	93.1
檮原町	77.5	84.7	85.3	76.1	83.0	82.9
大野見村	75.7	87.9	89.0	102.9	97.2	100.8
東津野村	112.9	106.6	108.6	98.4	95.8	98.5
葉山村	99.7	99.5	100.8	117.9	108.1	112.1
仁淀村	102.2	101.7	102.6	86.7	91.8	91.6
日高村	101.4	101.4	101.9	102.3	99.8	101.2
幡多郡						
佐賀町	122.1	108.9	114.6	128.5	112.0	117.5
大正町	82.6	89.9	90.9	88.2	90.7	92.2
大方町	103.1	102.1	103.1	107.9	104.9	106.2
大月町	114.2	108.0	111.4	96.5	97.5	97.0
十和村	99.4	99.4	100.8	100.8	97.2	100.0
西土佐村	92.8	97.1	96.5	106.6	102.4	103.7
三原村	104.0	101.3	103.5	87.8	96.0	93.8

注　ベイズ推定値では二次医療圏ごと，経験ベイズ推定値では県全体でベイズ推定を行っている。

（３）　市区町村別生命表

x 歳以上 $x+n$ 歳未満の中央死亡率 ${}_nm_x$ の事前分布を，密度関数が

$$\frac{1}{B(\alpha,\ \beta)}{}_nm_x^{\alpha-1}(1-{}_nm_x)^{\beta-1}$$

であるようなベータ分布であるとし，事後分布の期待値（平均）が，死亡数を D，人口を P としたとき，

$$\frac{\alpha+D}{\alpha+\beta+P}$$

で表される。

（４）　経験ベイズ推定法を用いた市区町村別標準化死亡比

ここでは，丹後俊郎・今井淳監修の Disease mapping system（疾病地図描画）を用いて，emprical Bayes 法を用いた市区町村別標準化死亡比（経験ベイズ推定値）を算出した。使用したデータは，平成10～14年の高知県の実績である。経験ベイズ（emprical Bayes）の方法により，次式が導かれる。

（標準化死亡比の経験ベイズ推定値）
＝w×(当該市区町村の標準化死亡比)
＋(1－w)×(県全体の標準化死亡比の平均)

ここに w は重みで，人口が増加すると１に近づき，減少すると０に近づく。したがって，人口が極めて小さい市区町村の標準化死亡比の経験ベイズ推定値は県全体の標準化死亡比の平均となり，反対に極めて人口が大きい市区町村の標準化死亡比の経験ベイズ推定値は当該市区町村の標準化死亡比そのものとなる性質を有している。表１が計算結果であり，図２から図５がそれを地図に表したもので

ある。地図をみると，標準化死亡比の経験ベイズ推定値では，120以上の市区町村と80未満の市区町村が標準化死亡比よりかなり減少していることがわかる。

図２　高知県　標準化死亡比（男，平成10～14年）

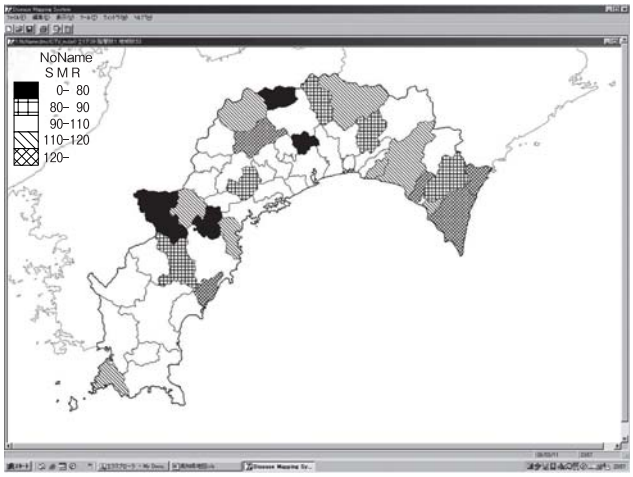
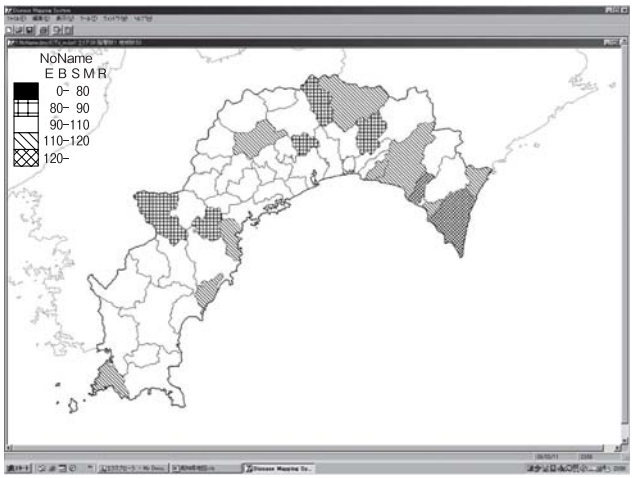


図３　高知県　標準化死亡比（男，平成10～14年，経験ベイズ推定値（経験ベイズ推定法））



Ⅴ　課題と展望

人口動態統計のような全数統計から指標を算出するとき，

・母集団の大きさに比して発生率が極めて小さい場合

・特に，母集団が小規模の場合

には，死亡を確率事象と考えると，全数統計であっても偶然変動を含むため，事象の発生件数から算出される頻度論的確率では，想定される真の確率からの乖離が大きくなりうる。この乖離を，標本抽出による標本誤差と区別して，仮に「モデル誤差」と呼ぶこととしたが，このモデル誤差を処理するには，

①　想定される値を設定する（事前分布）

②　実績により想定値を補整計算して真の値を得る（事後分布）

という論理（ベイズ推定）が有効である。市町村の指標を得ようとするときのポイントは，以下の通りである。

①　想定される値の設定(事前分布の設定)

具体的には，データの特性により，地域（保健所単位，二次医療圏，都道府県，地域ブロック，国全体等）と，求める統計のデータ分布形を選択する。

②　ベイズの定理により，得られた実績から想定値を補整計算する。

これまでのベイズ推定では，補整のための積分計算の数学的な難しさから，分布形を二項分布，ポアソン分布，正規分布，指数分布等の，数学的性質がよく知られたものにしなければいけないという制約があった。

一方，実務的には２次元以上の分布や経験的な分布も使いたいという要請がある。また，地域の選定についても，何を判断基準とするか議論があるところである。

コンピューターの計算速度の向上により，厳密な数学の計算はできなくとも近似的なシミュレーションにより補整計算を行うことのできる方法（MCMC 法）が開発された。シミュレーションであるので，事前分布の形の制約は受けず，ベイズ統計学の利用範囲が拡大する。また，計算が速いので，様々な地域で設定して，比較分析することも容易である。一方，シミュレーションによる有限計算なので，厳密な数学の計算と違い，計算上の誤差や，初期値の設定に

よっては本来の値とは違う値を得てしまう可能性もあり，シミュレーションの結果の妥当性には注意が必要である。

本論文では，平成10～14年の高知県の死亡実績をもとに，市区町村別標準化死亡比とそのベイズ推定値（二次医療圏）および経験ベイズ推定値（県全体）の比較を行い，小地域間の偶然変動によるばらつきをかなり小さくできるという結果を得た。また，さらに詳細な比較分析を行い，地域選定の判断基準を得たかったが，そこまでは至らなかった。地域選定の考え方としては，

① 例えば地域医療計画や保健所の公衆衛生活動のように，行政政策の実施単位地域とする考え方（最近の医療制度改革案では，都道府県単位で医療費適正化等に取り組むので，都道府県単位という考え方もあり得る）

② 統計的手法（分散分析など）により，着目するデータの選定地域間格差と地域内格差を比較して設定する考え方

③ ②を発展させて，文化や習慣など着目するデータ以外の指標を用いて，均質と考えられる地域単位を設定する考え方

が考えられるが，広い地域を設定すればするほど得られる統計データは安定するが，地域特性の反映度はますます薄くなるので，行政的に意味のある結果が得られるよう設定することが重要である。今後の課題として，上記の地域選定の判断基準の分析と，新たな指標を設定し，ベイズ推定法等による計算とその妥当性について研究することが挙げられる。

図４ 高知県 標準化死亡比（女，平成10～14年）

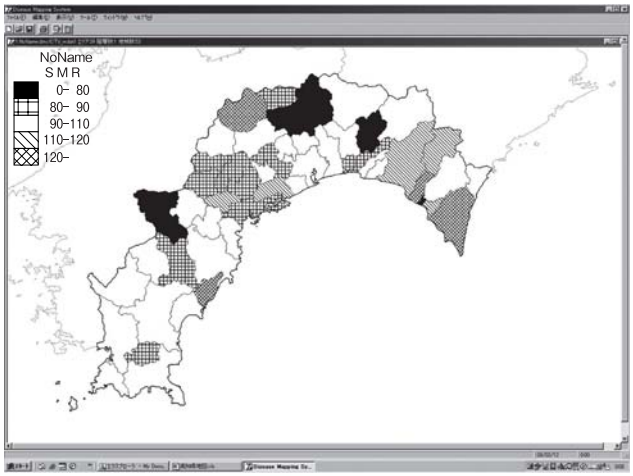
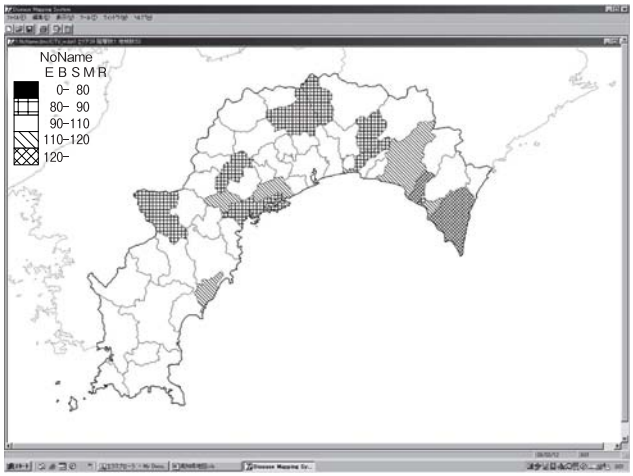


図５ 高知県 標準化死亡比（女，平成10～14年，経験ベイズ推定値（経験ベイズ推定法））



文　献

1）丹後俊郎．統計モデル入門．東京：朝倉書店，2000.

2）汪金芳，田栗正章，手塚集，他．計算統計Ⅰ 確率計算の新しい手法．甘利俊一，竹内啓，竹村彰通，他編．統計科学のフロンティア11．東京：岩波書店，2003.

3）伊庭幸人，種村正美，大森裕浩，他．計算統計Ⅱ マルコフ連鎖モンテカルロ法とその周辺．甘利俊一，竹内啓，竹村彰通，他編．統計科学のフロンティア12．東京：岩波書店，2005.

4）下平英寿，伊藤秀一，久保川達也，他．モデル選択 予測・検定・推定の交差点．甘利俊一，竹内啓，竹村彰通，他編．統計科学のフロンティア3．東京：岩波書店，2004.

5）石黒真木夫，松本隆，乾敏郎，他．階層ベイズモデルとその周辺 時系列・画像・認知への応用．甘利俊一，竹内啓，竹村彰通，他編．統計科学のフロンティア4．東京：岩波書店，2004.