ベイズ統計学 (基礎理論篇)

渡部元博

1/23/23

Table of contents

1 用語の整理

共役性 θ に対する事前分布のクラス $\mathcal{P}$ が標本モデル $p(y|\theta)$ に対して共役であるとは、

$$P(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

となることをいう。

共役事前分布は事後計算を容易にするが、実際には事前情報を表してない場合がある。ただし、共役事前分布の混合分布は非常に柔軟であり、計算上扱いやすいものである。

情報の統合 ベータ分布を例にする。$\theta|\{Y=y\} \sim beta(a+y, b+n-1)$ なら、

$$E[\theta|y] = \frac{a+y}{a+b+n}, \quad mode[\theta|y] = \frac{a+y-1}{a+b+n-2}, \quad Var[\theta|y] = \frac{E[\theta|y]E[1-\theta|y]}{a+b+n+1}$$

となる。

事後期待値 $E[\theta|y]$ は事前の情報とデータの情報の統合であることが次の式から容易にわかる。

$$\begin{aligned} E[\theta|y] &= \frac{a+y}{a+b+n} \\ &= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\ &= \frac{a+b}{a+b+n} \times PreExpectant + \frac{n}{a+b+n} \times mean \end{aligned}$$

このモデルと事前分布に対して、事後期待値は、事前期待値と標本平均の加重平均であり、重みはそれぞれ a+b と n に比例する。これにより、a と b が「事前のデータ」として解釈される。

予測　二値の場合を考えると、$\tilde{Y} \in \{0,1\}$ を同じ母集団からのまだ観測していない確率変数とする。$\tilde{Y}$ の**予測分布**は、$\{Y_1 = y_1, \cdots, Y_n = y_n\}$ が与えられたもとでの $\tilde{Y}$ の条件付き分布である。条件付き独立同一の二値確率変数に対して、この予測分布は θ を与えたもとでの $\tilde{Y}$ の分布と θ の事後分布から導かれる。

$$\begin{aligned}
Pr(\tilde{Y} = 1|y_1, \cdots, y_n) &= \int Pr(\tilde{Y} = 1, \theta|y_1, \cdots, y_n)d\theta \\
&= \int Pr(\tilde{Y} = 1, |\theta, y_1, \cdots, y_n)p(\theta|y_1, \cdots, y_n)d\theta \\
&= \int \theta p(\theta|y_1, \cdots, y_n)d\theta \\
&= E[\theta|y_1, \cdots, y_n] = \frac{a + \sum_{i=1}^{n} y_i}{a + b + n}, \\
Pr(\tilde{Y} = 0|y_1, \cdots, y_n) &= 1 - E[\theta|y_1, \cdots, y_n] = \frac{b + \sum_{i=1}^{n}(1 - y_i)}{a + b + n}
\end{aligned}$$

予測分布については次の二つの重要な点に注意する必要がある。

1.　予測分布は、未知の量に依存しない。もし、それが未知の量に依存しているなら、それを使って予測することはできない。

2.　予測分布は観測データに依存する。この予測分布において、$\tilde{Y}$ は $Y_1, \cdots, Y_n$ とは独立ではない。これは、$Y_1, \cdots, Y_n$ を観測することにより θ に関する情報が得られ、それが $\tilde{Y}$ に関する情報を与えるからである。もし、$\tilde{Y}$ が $Y_1, \cdots, Y_n$ とは独立であるならば、それはよくないことである。それは対象としている母集団からサンプリングされていない量について何も推測できないことを意味するからである。

ベイズ信用区間　観測データ $Y{=}y$ に基づく区間 $[l(y), u(y)]$ が、θ に対する 95% **信用区間**であるとは、

$$Pr(l(y) < \theta < u(y)|Y = y) = 0.95$$

が成り立つことをいう。

これは、$Y{=}y$ を観測した後に、θ の真の値がどの位置にあるかという情報を表す区間であり、データが観測される「前に」区間が真の値を被覆する確率を説明するような頻度論的解釈とは異なる。

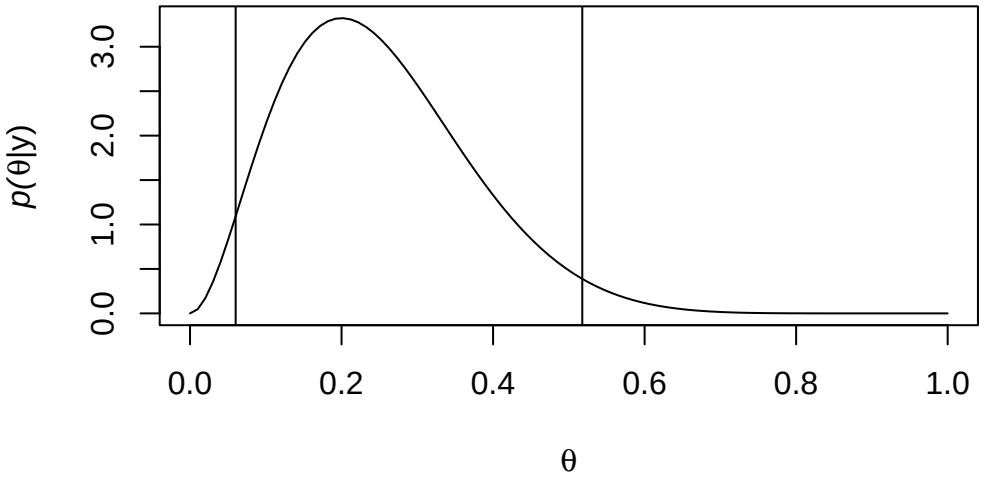
頻度論的信頼区間　ランダムな区間 $[l(Y), u(Y)]$ が、θ に対する 95% 信頼区間であるとは、データが得られる前に

$$Pr(l(Y) < \theta < u(Y)|\theta) = 0.95$$

が成り立つことをいう。

信用区間と信頼区間の概念は、それぞれ実験前と実験後に構成される区間としての違いがある。

単純な 95% 信用区間では推測のバイアスが大きくなってしまうことがある。以下の図はその一例である。

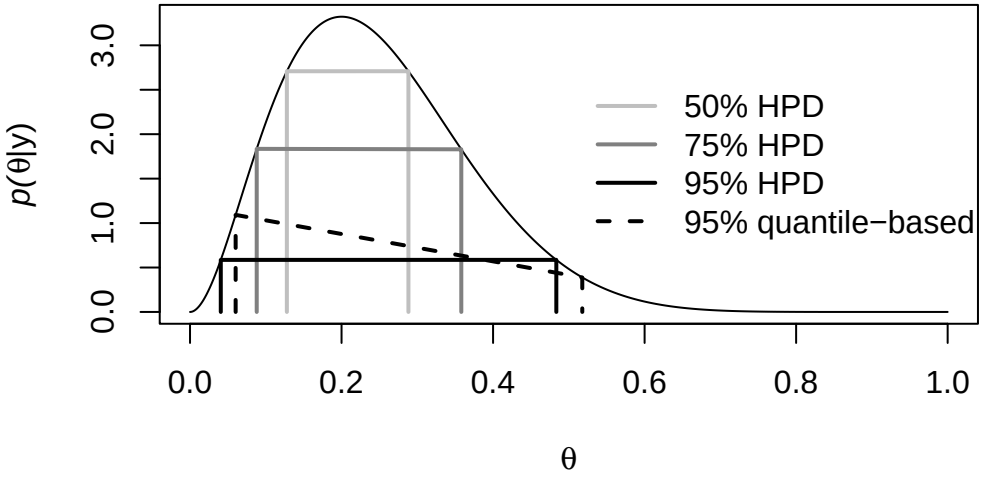


分位点に基づく区間の外側に、区間内のいくつかの点よりも高い確率 (密度) をもつ θ の値があることに注意する。これは、より制約のある区間の存在を示している。

最高事後密度 (Highest Posterior Density, HPD) 領域　$100(1-\alpha)\%$HPD 領域は次を満たすパラメータ空間の部分集合 $s(y) \subset \Theta$ で構成される。

1. $Pr(\theta \in s(y)|Y = y) = 1 - \alpha$
2. $\theta_\alpha \in s(y)$ かつ $\theta_\beta \notin s(y)$ ならば、$p(\theta_\alpha|Y = y) > p(\theta_\beta|Y = y)$

HPD 領域内の全ての点では、領域外の点よりも事後密度が高くなる。ただし、事後密度が多峰型ならば、HPD 領域は区間ではない可能性がある。ここからは視覚的に HPD 領域の構築方法について確認していく。



まず、水平線を密度全体で徐々に下に移動する。これには、HPD 領域で、水平線より上の密度を持つ全ての θ の値が含まれる。領域内の θ の事後確率が $(1-\alpha)$ に達したら、線を下に移動するのを停止する。上図の場合、95%HPD 領域は [0.04,0.48] であり、分位点に基づく区間よりも狭いが、両方とも事後確率の 95% を含んでいる。

2　ベイズの哲学

標準ベイズ 2 章 (8.2.1 も)

3　確率論

4　確率分布論

確率分布については、Julia のパッケージである Pluto でインタラクティブなノートブックを作成しているので、そちらも参照されたい。はじめにベイズ統計の基礎を概観するために、一つの未知パラメータにより定まる確率分布である二項モデルとポアソンモデルに対するベイズ推測について考える。

4.1　二項モデル

データは、1998 年の総合的社会調査において、65 歳以上の女性に総じて幸せかどうか質問する調査が行われた。もし、回答者 i が総じて幸せであると回答したら、$Y_i = 1$ とし、そうでなければ $Y_i = 0$ と置くことにする。もし、彼女ら 129 人を区別する情報が欠如している場合は、これらの回答は交換可能であるものとして良いと考えられる。129 人というのは、高齢の女性の全数 N よりもはるかに少ないので、各 Y の同時信念は以下のように記述できる。

- $\theta = \sum_{i=1}^{N} Y_i/N$ に関する信念
- θ で条件づけされたもとで、各 Y は期待値 θ を持つ独立同一の二値の確率変数であるというモデル

後者は θ で条件づけされたもとで、任意の潜在的な回答$\{y_i, \cdots, y_{129}\}$に対する確率が

$$p(y_1, \cdots, y_{129}|\theta) = \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i}$$

で与えられることを意味している。

次に事前分布を定義する。パラメータ θ はある未知の数で、0 から 1 の間に値を取る。θ に関する事前情報としては、同じ区間幅を持つ [0,1] の全ての部分区間において同じ確率をもつと仮定しよう。これを数式で書くと以下のようになる。

$$Pr(a \leq \theta \leq b) = Pr(a + c \leq \theta \leq b + c) \quad (0 \leq< a < b < b + c \leq 1)$$

この条件により、θ に関する密度関数は一様分布の密度関数にならないといけないことがわかる。つまり、

$$\forall\theta \in [0, 1] \quad p(\theta) = 1$$

この事前分布と、上記の標本モデルに対して、ベイズルールにより以下が成り立つ。

$$\begin{aligned} p(\theta|y1,\cdots,y_{129}) &= \frac{p(y1,\cdots,y_{129}|\theta)p(\theta)}{p(y1,\cdots,y_{129})} \\ &= p(y1,\cdots,y_{129}|\theta) \times \frac{1}{p(y1,\cdots,y_{129})} \\ &\propto p(y1,\cdots,y_{129}|\theta) \end{aligned}$$

最後の式は、$p(\theta|y1,\cdots,y_{129})$ と $p(y1,\cdots,y_{129}|\theta)$ は θ の関数として比例の関係にあることを意味している。というのも、事後分布が $p(y1,\cdots,y_{129}|\theta)$ を θ に依存しないもので割ったものに等しいからである。このことから、これら二つの θ の関数は同じ形状を持つが、必ずしも同じ尺度をもつとは限らないことがわかる。

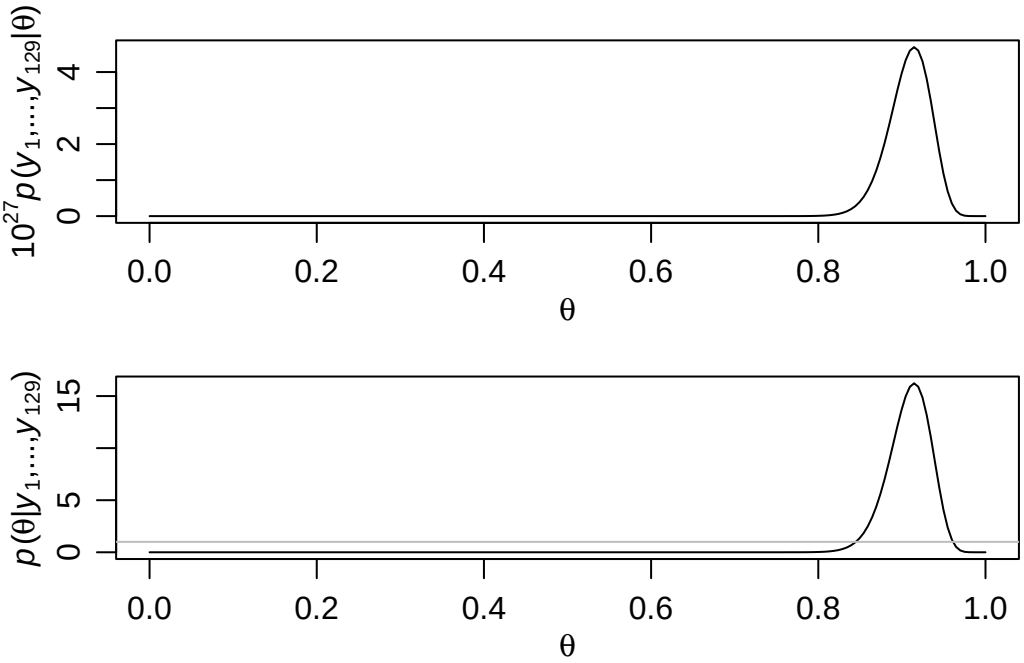
データの詳細は以下 3 点である。

- 調査対象は 129 人
- 総じて幸せであると回答したのは 118 人 (91%)
- 総じて幸せであると回答しなかったのは 11 人 (9%)

ある θ の値が与えられたもとでのこれらのデータの確率は

$$p(y1,\cdots,y_{129}|\theta) = \theta^{118}(1-\theta)^{11}$$

となる。この確率を θ の関数とみて図示したものが以下の図 (上段) である。



事後分布 $p(\theta|y1,\cdots,y_{129})$ はこれと同じ形状をしていて、θ の真の値は 0.91 付近にあり、0.80 より大きいことはほぼ確実であることを示唆している。しかし、事後分布を正確に計算したいことがよくあり、そのためには、$p(\theta|y1,\cdots,y_n)$ の形状に加え、尺度も知る必要があるであろう。ベイズルールにより、以下が成り立つ。

$$\begin{aligned} p(\theta|y1,\cdots,y_{129}) &= \theta^{118}(1-\theta)^{11} \times p(\theta)/p(y_1,\cdots,y_{129}) \\ &= \theta^{118}(1-\theta)^1 1 \times 1/p(y_1,\cdots,y_{129}) \end{aligned}$$

また、ベータ関数とガンマ関数の関係

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

を用いて、尺度あるいは**正規化定数** $1/p(y_1,\cdots,y_{129}$ を計算することができる。この関係がどのように $p(y_1,\cdots,y_{129})$ の計算に使われるのだろうか。まず、$p(\theta|y_1,\cdots,y_{129})$ に関してわかっていることを以下にまとめる。

1. 全ての確率分布は全積分もしくは総和が 1 になる
2. ベイズルールにより、$p(\theta|y1,\cdots,y_{129}) = \theta^{118}(1-\theta)^{11}/p(y_1,\cdots,y_{129})$

ゆえに、

$$\begin{aligned}
1 &= \int_0^1 p(\theta|y_1,\cdots,y_{129})d\theta \\
1 &= \int_0^1 \theta^{118}(1-\theta)^{11}/p(y_1,\cdots,y_{129})d\theta \\
1 &= \frac{1}{p(y_1,\cdots,y_{129})}\int_0^1 \theta^{118}(1-\theta)^{11}d\theta \\
1 &= \frac{1}{p(y_1,\cdots,y_{129})}\frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}
\end{aligned}$$

となるため、

$$p(y_1,\cdots,y_{129}) = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}$$

となる。これら全てをまとめると、

$$p(\theta|y_1,\cdots,y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{119-1}(1-\theta)^{12-1}$$

となる。この θ に関する密度は、パラメータ a = 119 と b=12 をもつ**ベータ分布**と呼ばれる。

このベータ分布にに従う確率変数は

$$\begin{aligned}
mode[\theta] &= (a-1)/[(a-1)+(b-1)] \quad (a,b>1) \\
E[\theta] &= a/(a+b) \\
Var[\theta] &= ab/[(a+b+1)(a+b)^2] = E[\theta]\times E[1-\theta]/(a+b+1)
\end{aligned}$$

が成り立つ。今回の幸福度データの場合、最頻値が 0.915、平均値が 0.908、分散が 0.025 となる。

詳細は割愛するが、Y が二値変数である時の未知パラメータ θ と事前分布 $p(y_1,\cdots,y_{129}|\theta)$ の十分統計量は $\sum_i Y_i$ となり、パラメータ (n,θ) の**二項分布**に従う。ここで、確率変数 $Y \in \{0,1,\cdots,n\}$ が、

$$Pr(Y=y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

となることをいう。また、次の尺度を満たす。

$$\begin{aligned}
E[Y|\theta] &= n\theta \\
Var[Y|\theta] &= n\theta(1-\theta)
\end{aligned}$$

4.1.1 二項モデルの事後推測

$Y = y$ を観測した時、θ の事後分布を求めるためには、以下の計算をする。

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{p(y)} \\ &= c(y)\theta^y(1-\theta)^{n-y}p(\theta) \end{aligned}$$

ここで、$c(y)$ は θ に依存していない y の関数である。一様分布 $p(\theta) = 1$ に対して、微積分の計算をすることで、具体的な $c(y)$ を求めることができる。

$$\begin{aligned} 1 = \int_0^1 c(y)\theta^y(1-\theta)^{n-y}d\theta &\iff 1 = c(y)\int_0^1 \theta^y(1-\theta)^{n-y}d\theta\frac{p(y|\theta)p(\theta)}{p(y)} \\ &\iff 1 = c(y)\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \end{aligned}$$

よって、正規化定数 $c(y)$ は上記式二段目の分数の逆数となる。よって、一様分布が与えられた際の事後分布は、以下のように書き下すことができる。

$$\begin{aligned} p(\theta|y) &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1} \\ &= beta(y+1, n-y+1) \end{aligned}$$

幸福度の場合、一様分布を事前分布として与えた事後分布は、

$$n = 129, Y \equiv \sum Y_i = 118, \quad \Rightarrow \quad \theta|\{Y = 118\} \sim beta(119, 12)$$

次に、事前分布としてベータ分布を与えた事後分布の推測を行う。先述の一様事前分布は a=1,b=1 をもつベータ分布として見ることができる。以下に挙げるガンマ関数の特徴を用いて証明ができる。 - $\Gamma(x+1) = x! = n\Gamma(n)$ - $\Gamma(1) = 1$

$\theta \sim beta(a, b)$ とし、$Y|\theta \sim binomial(n, \theta)$ とする。Y=y を観測した時、

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{1}{p(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y}\theta^y(1-\theta)^{n-y} \\ &= c(n.y, a, b) \times \theta^{a+y-1}(1-\theta)^{b+n+y-1} \\ &= beta(a+y, b+n-y) \end{aligned}$$

式三段目から、事後分布がベータ分布と同じ「形状」を有することがわかる。また、両者は積分して 1 にならないいけないので、同じ「尺度」も共有する。このことから事後分布とベータ密度が実際には同じ関数である

ことを意味していることが窺える。ベイズ統計学では、このような方法を通して事後分布を識別していく。つまり、事後分布は既知の確率密度に比例するため、その密度と等しくなければならない。今回の例のようにベータ事前分布を与えることで二項モデルの事後分布を推測できた。このとき、ベータ事前分布は**共役性**を有する。

4.2 ポアソンモデル

4.2.1 ポアソン分布

確率変数 Y が平均 θ のポアソン分布に従うとは、

$$Pr(Y = y|\theta) = dpois(y, \theta) = \frac{\theta^y e^{-\theta}}{y!} \quad y \in \{0, 1, 2, \cdots\}$$

となることである。そのとき、尺度は

* $E[Y|\theta] = \theta$
* $Var[Y|\theta] = \theta$

が成り立つ。ポアソン分布の平均が大きい場合、その分散も大きくなるため、ポアソン分布族には「平均分散関係」があるともいう。

4.2.2 ポアソンモデルの事後推測

$Y_1, \cdots, Y_n$ を平均 θ のポアソン分布からの独立同一標本であるとモデル化すると、同時確率密度は、

$$\begin{aligned} Pr(Y_1 = y_1, \cdots, Y_n = y_n|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{y!}\theta^{y_i} e^{-\theta} \\ &= c(y_1, \cdots, y_n)\theta^{\sum y_i} e^{-n\theta} \end{aligned}$$

となる。十分統計量は $\sum_{i=1}^{n} Y_i$ で、十分統計量は平均 $n\theta$ のポアソン分布に従う。

ポアソン分布の事後分布は次のようになる。

$$\begin{aligned} p(\theta|y_1, \cdots, y_n) &\propto p(\theta) \times p(y_1, \cdots, y_n|\theta) \\ &\propto p(\theta) \times \theta^{\sum y_i} e^{-n\theta} \end{aligned}$$

これは、共役分布族の密度が何であれ、c_1 と c_2 に対して $\theta^{c_1} e^{-c_2\theta}$ のような項を含める必要があることを意味している。このような密度の最も単純なクラスに対応する確率分布として、これらの項のみが含まれるガンマ分布が知られている。ガンマ分布の形状は

$$p(\theta) \equiv dgamma(\theta, a, b) = \frac{b^n}{\Gamma(a)}\theta^{a-1} e^{-b\theta} \quad (\theta, a, b > 0)$$

であり、尺度は以下を満たす。

* $E[\theta] = a/b$

* $Var[\theta] = a/b^2$

* $mode[\theta] = (a-1)b \quad (a > 1)$

$Y_1, \cdots, Y_n|\theta \sim i.i.d.Poisson(\theta), p(\theta) = dgamma(\theta, a, b)$ とする。このとき、事後分布は、

$$\begin{aligned} p(\theta|y_1, \cdots, y_n) &= p(\theta) \times p(y_1, \cdots, y_n|\theta)/p(y_1, \cdots, y_n) \\ &= \{\theta^{a-1}e^{-b\theta}\} \times \{\theta^{\sum y_i}e^{-n\theta}\} \times c(y_1, \cdots, y_n, a, b) \\ &= \{\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\} \times c(y_1, \cdots, y_n, a, b) \end{aligned}$$

となる。これは明らかにガンマ分布であり、ガンマ分布族のポアソンモデルに対する共役性が確認できる。

$$\left.\begin{array}{r} \theta \sim gamma(a, b) \\ Y_1, \cdots, Y_n|\theta \sim Poisson(\theta) \end{array}\right\} \Rightarrow \{\theta|Y_1, \cdots, Y_n\} \sim gamma(a + \sum_{i=1}^{n} Y_i, b + n)$$

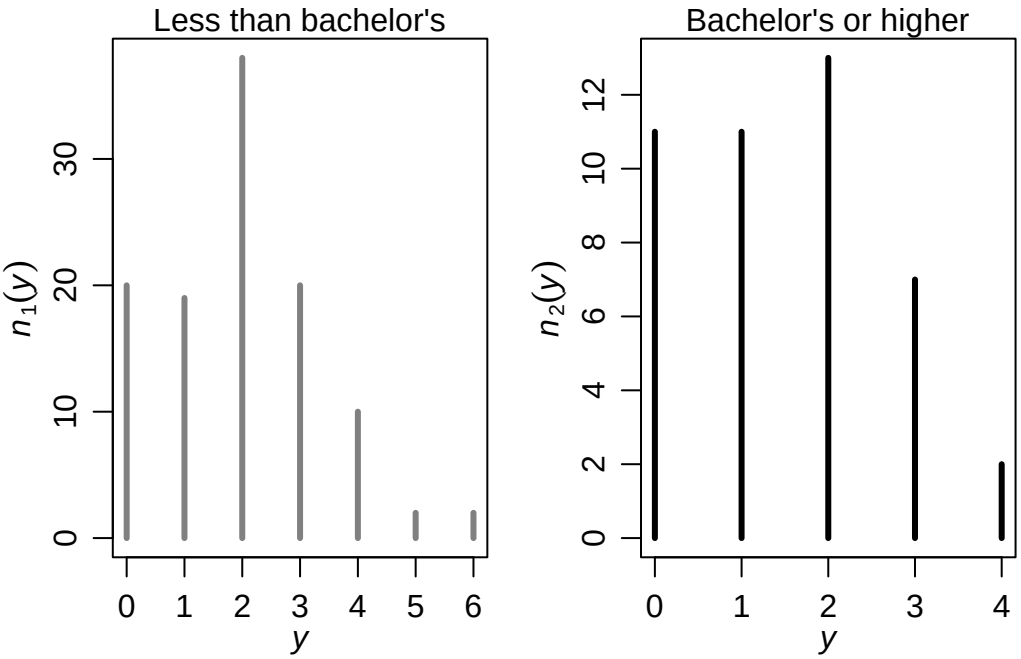
推定と予測は、二項モデルと同様の方法で行うことができる。θ の事後期待値は、事前期待値と標本平均の凸結合になる。

4.2.3 例: 出生率

1990 年台に、総合的社会調査は、調査への参加時に 40 歳であった 155 人の女性の学歴と子供の数に関するデータを収集した。これらの女性は、歴史的に見ても出生率が低かった 1970 年代の 20 代に相当する。この例では、学士号を持つ女性とそうでない女性の子供の数を比較する。$Y_{1,1}, \cdots, Y_{n1,1}$ を学士号を持たない n_1 人の女性の子供の数とし、$Y_{1,2}, \cdots, Y_{n2,2}$ を学士号を持つ n_2 人の女性の子供の数とする。この例では、次のようなサンプリングモデルを用いる。

$$Y_{1,1}, \cdots, Y_{n1,1} \overset{\text{iid}}{\sim} Poisson(\theta_1) Y_{1,2}, \cdots, Y_{n2,2} \overset{\text{iid}}{\sim} Poisson(\theta_2)$$

データの経験分布を以下に示す。



グループの合計と平均は次で与えられる。

$$less\ than\ Bachelor: n_1 = 111, \sum_{i=1}^{n_1} Y_{i,1} = 217, \bar{Y}_1 = 1.95, more\ than\ Bachelor: n_2 = 44, \sum_{i=1}^{n_2} Y_{i,2} = 66, \bar{Y}_2 = 1.50,$$

$\{\theta_1, \theta_2\} \overset{\text{iid}}{\sim} gamma(a = 2, b = 1)$ の場合、次の事後分布を得る。

$$\theta_1 | \{n_1 = 111, \sum Y_i, 1 = 217\} \sim gamma(2 + 217, 1 + 111) = gamma(219, 112)$$
$$\theta_2 | \{n_1 = 44, \sum Y_i, 1 = 66\} \sim gamma(2 + 66, 1 + 44) = gamma(68, 45)$$

θ_1 と θ_2 の事後平均、最頻値、分位点に基づく 95% 信用区間は、それらのガンマ事後分布から得られる。

```
a <- 2; b <- 1 #事前分布のパラメータ
n1 <- 111; sy1 <-217 #大卒未満のデータ
n2 <- 44; sy2 <- 66 #大卒以上のデータ

(a+sy1)/(b+n1) #事後平均
```

```
[1] 1.955357
```

```
(a+sy1-1/(b+n1)) #事後最頻値
```

```
[1] 218.9911
```

```
qgamma(c(0.0025,0.975),a+sy1,b+n1) #95% 事後信用区間
```

```
[1] 1.604883 2.222679
```

```
(a+sy2)/(b+n2) #事後平均
```

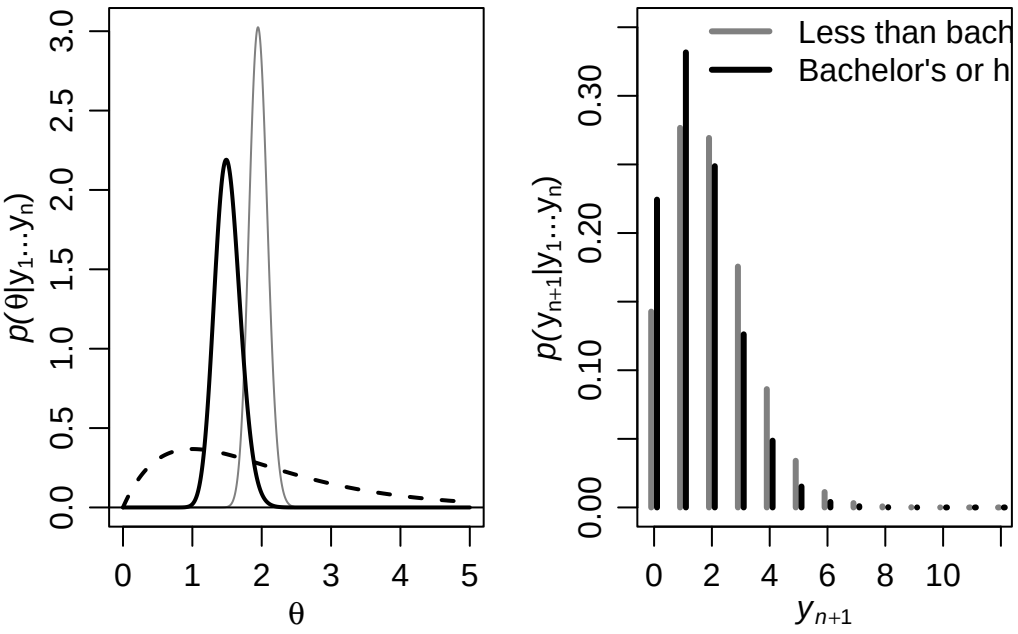
```
[1] 1.511111
```

```
(a+sy2-1/(b+n2)) #事後最頻値
```

```
[1] 67.97778
```

```
qgamma(c(0.0025,0.975),a+sy2,b+n2) #95% 事後信用区間
```

```
[1] 1.047401 1.890836
```



二つのグループの母平均の事後密度は、上の図左に示されている。事後分布は、実質的に $\theta_1 > \theta_2$ と言えるほどの証拠を示している。実際に $Pr(\theta_1 > \theta_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = 0.97$ である。ここで、二つの母集団からそれぞれ一人ずつ、ランダムにサンプリングされた 2 人について考える。学士号を取得していない方が、もう一方よりも子供が多いとどの程度期待されるだろうか。$\tilde{Y}_1$ と $\tilde{Y}_2$ に対する事後予測分布はともに負の二項分布になり、図の右側である。θ_1 と θ_2 の事後分布の間の差を考えた時よりも、二つの予測分布の間にはより多くの重複があることに気づくだろう。たとえば、$Pr(\tilde{Y}_1 > \tilde{Y}_2) | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = 0.48$ となる。事象 $\{\theta_1 > \theta_2\}$ と $\{\tilde{Y}_1 > \tilde{Y}_2\}$ の区別は非常に重要である。二つの母集団に差があるという強い証拠があるからといって、その差が大きいとは限らない。

4.3 単変量正規モデル

確率変数 Y が平均 θ、分散 $\sigma^2 > 0$ の正規分布に従うとは、Y の密度関数が以下で与えられる時をいう。

$$p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\theta}{\sigma})^2}, \quad -\infty < Y < \infty$$

以下に特筆すべき点を挙げる。

* この分布は θ に関して対称であり、平均、中央値、最頻値は全て θ に等しい。

* 分布のおよそ 95% は平均から標準偏差の 2 倍の距離の間にある (より正確には 1.96 倍)

$X \sim normal(\mu,\tau^2), \quad Y \sim normal(\theta,\sigma^2)$ かつ X と Y*が独立ならば、$aX + bY \sim normal(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)$ となる。

正規分布の重要性は、主に中心極限定理に由来する。この定理は非常に一般的な状況的なもとで、確率変数の和 (または平均) は近似的に正規分布に従うと主張する。多くの要素を足してできるものから生じるデータに対して、正規分布に基づく標本モデルは適切だということになる。 ### 分散既知での推測

モデルとして、$\{Y_1,\cdots,Y_n|\theta,\sigma^2\} \overset{\text{iid}}{\sim} normal(\theta,\sigma^2)$ を考えよう。同時密度関数は以下のようになる。

$$\begin{aligned} p(y_1,\cdots,y_n|\theta,\sigma^2) &= \prod_{i=1}^{n} p(y_i|\theta,\sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i-\theta}{\sigma})^2} \\ &= (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2}\sum(\frac{y_i-\theta}{\sigma})^2\right\} \end{aligned}$$

指数の中にある 2 次の項を展開すると、$p(y_1,\cdots,y_n|\theta,\sigma^2)$ が $y_1,\cdots,y_n$ にどう依存しているかが、以下のようにしてわかる。

$$\sum_{i=1}^{n}\left(\frac{y_i-\theta}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum y_i^2 - 2\frac{\theta}{\sigma^2}\sum y_i + n\frac{\theta^2}{\sigma^2}$$

これより、$\{\sum y_i^2, \sum y_i\}$ が二次元の十分統計量になることが示される。これらの値を知ることは、$\bar{y} = \sum y_i/n$ と $s^2 = \sum(y_i-\bar{y})^2/(n-1)$ の値を知ることと同値であるので、$\{\bar{y}, s^2\}$ もまた十分統計量である。

二つのパラメータをもつモデルに関する推測は、一つのパラメータに関する二つの問題に分割される。まず、σ^2 が吉の場合の θ の推測に関する問題に取り組み、θ の共役事前分布を使用する。任意の (条件付き) 事前分布 $p(\theta|\sigma^2)$ について、事後分布は以下を満たす。

$$\begin{aligned} p(\theta|y_1,\cdots,y_n,\sigma^2) &\propto p(\theta|\sigma^2) \times e^{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2} \\ &\propto p(\theta|\sigma^2) \times e^{c_1(\theta-c_2)^2} \end{aligned}$$

上の計算により、$p(\theta|\sigma^2)$ が共役になるためには、それは $e^{c_1(\theta-c_2)^2}$ のような二次の項を含んでいなければならない。そのような $\mathbb{R}$ 上の確率分布のグラフで最も単純なものは正規分布族である。実際に確かめてみる。

$\theta \sim normal(\mu_0, \tau_0^2)$ の時、

$$\begin{aligned} p(\theta|y_1, \cdots, y_n, \sigma^2) &\propto p(\theta|\sigma^2)p(y_1, \cdots, y_n|\theta, \sigma^2)/p(y_1, \cdots, y_n|\sigma^2) \\ &\propto p(\theta|\sigma^2)p(y_1, \cdots, y_n|\theta, \sigma^2) \\ &\propto exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\} exp\left\{-\frac{1}{2\sigma_0^2}\sum(y_i - \theta)^2\right\} \end{aligned}$$

となる。指数の中にある和を計算し、-1/2 をひとまず無視することで、岡の表現を得る。

$$\frac{1}{\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}\left(\sum y_i^2 - 2\theta\sum y_i + n\theta^2\right) = a\theta^2 - 2b\theta + c, a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}, \quad b = \frac{\mu_o}{\tau_0^2} + \frac{\sum y_i}{\sigma^2}, \quad c = c(\mu_0, \tau_0^2, \sigma_0^2, y_1, \cdots, y_r$$

さて、$p(\theta|\sigma^2, y_1, \cdots, y_n)$ が正規分布の密度関数の形になることを確かめよう。

$$\begin{aligned} p(\theta|\sigma^2, y_1, \cdots, y_n) &\propto exp\left\{-\frac{1}{2}(a\theta^2 - 2b\theta)\right\} \\ &= exp\left\{-\frac{1}{2}a(\theta^2 - 2b\theta/a + b^2/a^2) + \frac{1}{2}b^2/a\right\} \\ &\propto exp\left\{-\frac{1}{2}a(\theta - b/a)^2\right\} \\ &= exp\left\{-\frac{1}{2}\left(\frac{\theta - b/a}{1/\sqrt{a}}\right)^2\right\} \end{aligned}$$

この関数の形状はまさに正規分布の密度と同じであり、$1/\sqrt{a}$ は標準偏差、b/a は平均の役割を果たしている。この分布の平均と分散をそれぞれ μ_n と τ_n^2 で表す。

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \mu_n = \frac{a}{b} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

4.3.1 分散未知での推測

θ と σ^2 に関する任意の事前分布 $p(\theta, \sigma^2)$ に対して、事後推測はベイズルール wp 用いて以下のように進行する。

$$p(\theta, \sigma^2|y_1, \cdots, y_n) = p(y_1, \cdots, y_n|\theta, \sigma^2)p(\theta, \sigma^2)/p(y_1, \cdots, y_n)$$

前節と同様に事後分布の計算を簡単にするため、単純で共役な事前分布のクラスを開発することから確かめよう。二つの変数に関する同時分布は条件付き確率と周辺確率の積で表せるという。以下の確率の公理を用いる。

$$p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$$

もし、σ^2 が既知ならば、θ の共役事前分布は $normal(\mu_0, \tau_0^2)$ である。特に、$\tau_0^2 = \sigma^2/\kappa_0$ という場合を考察しよう。

$$p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2) = dnorm(\theta, \mu_0, \tau_0 = \sigma/\sqrt{\kappa}_0 \times p(\sigma^2)$$

この場合、パラメータ μ_0 と κ_0 は、事前の観測値から得られた平均とサンプルサイズであると解釈できる。

σ^2 については、台が、$(0, \infty)$ となる事前分布族が必要となるので、ガンマ分布族が適しているように感じる。しかし、それらの分布族は正規分布の分散に対して共役ではない。しかし、$1/\sigma^2$ については共役になる

ことが明らかになる。そのような事前分布を用いるとき、σ^2 は**逆ガンマ分布**に従うという。事前分布を以下のパラメータで特徴づける。

$$1/\sigma^2 \sim gamma(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$$

事前分布と標本モデルを整理すると以下のようになる。

$$\theta|\sigma^2 \sim normal(\mu_0, \sigma^2/\kappa_0),$$
$$Y_1, \cdots, Y_n|\theta, \sigma^2 \overset{\text{iid}}{\sim} normal(\theta, \sigma^2)$$

θ と σ^2 の事前分布を $p(\theta, \sigma^2) = p(\theta|\sigma^2)p(\sigma^2)$ と分解できるように、事後分布も同じように分解できる。

$$p(\theta, \sigma^2|y_1, \cdots, y_n) = p(\theta|\sigma^2, y_1, \cdots, y_n)p(\sigma^2)p(\sigma^2|y_1, \cdots, y_n)$$

データと σ^2 が所与のもとで、の θ の条件付き分布は前節の結果を用いて得られる。つまり、τ_0^2 に σ^2/κ_0 を代入することで、以下を得る。

$$\{\theta|y_1, \cdots, y_n, \sigma^2\} \sim normal(\mu_n, \sigma^2/\kappa_n), \kappa_n = \kappa_0 + n, \quad \mu_0 = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$$

これより、もし μ_0 が κ_0 個の事前の観測値の平均ならば、$E[\theta|y_1, \cdots, y_n, \sigma^2]$ は事前及び現在に得た観測値の平均であり、$Var[\theta|y_1, \cdots, y_n]$ は事前および現在に得た観測値の数で σ^2 を割ったものである。　σ^2 の事後分布は未知の値 θ について積分することで得られる。

$$\begin{aligned} p(\sigma^2|y_1, \cdots, y_n) &\propto p(\sigma^2)p(y_1, \cdots, y_n|\sigma^2) \\ &= p(\sigma^2)\int p(y_1, \cdots, y_n|\theta, \sigma^2)p(\theta|\sigma^2)d\theta \end{aligned}$$

計算の結果は次のようになる。

$$\begin{aligned} \{1/\sigma^2|y_1, \cdots, y_n\} &\propto gamma(\nu_n/2, \nu_n\sigma_n^2/2) \\ v_0 &= v_0 + n \\ \sigma_n^2 &= \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2] \end{aligned}$$

これらの式から、v_0 は事前の観測値のサンプルサイズ、σ_0^2\$は事前の標本分散であるという解釈が得られる。

4.3.2　例: ミッジ (羽虫) の羽長

調査対象の母集団の真の平均と標準偏差はそれぞれ 1.9mm と,0.01mm から大きくしているべきではない。つまり、$\mu_0 = 1.9, \quad \sigma_0^2 = 0.01$ となることが示唆されている。しかし、羽の長さに関してこの母集団は他の母集団とは異なるので、これらの推定値を中心にしつつも事前分布が弱くなるよう。$\kappa_0 = \nu_0 = 1$ と定める。

われわれが観測したデータの標本平均 h$\bar{y} = 1.804, \quad s^2 = 0.0169$ である。これらの値と事前分布のパラメータから、μ_n と σ_n^2 を次のように計算する。

```
#事前分布
mu0 <- 1.9;k0 <- 1
s20 <- 0.010;nu0 <-1

#データ
y <- c(1.64,1.70,1.72,1.74,1.82,1.82,1.90,2.08)
n <- length(y); ybar <- mean(y); s2 <-var(y)

#事後推測
kn<- k0+n; nun <- nu0+n
mun <-(k0*mu0 + n*ybar)/kn
s2n <- (nu0*s20 + (n-1)*s2 + k0*n*(ybar-mu0)^2/(kn))/(nun)

nun
```

```
[1] 9
```

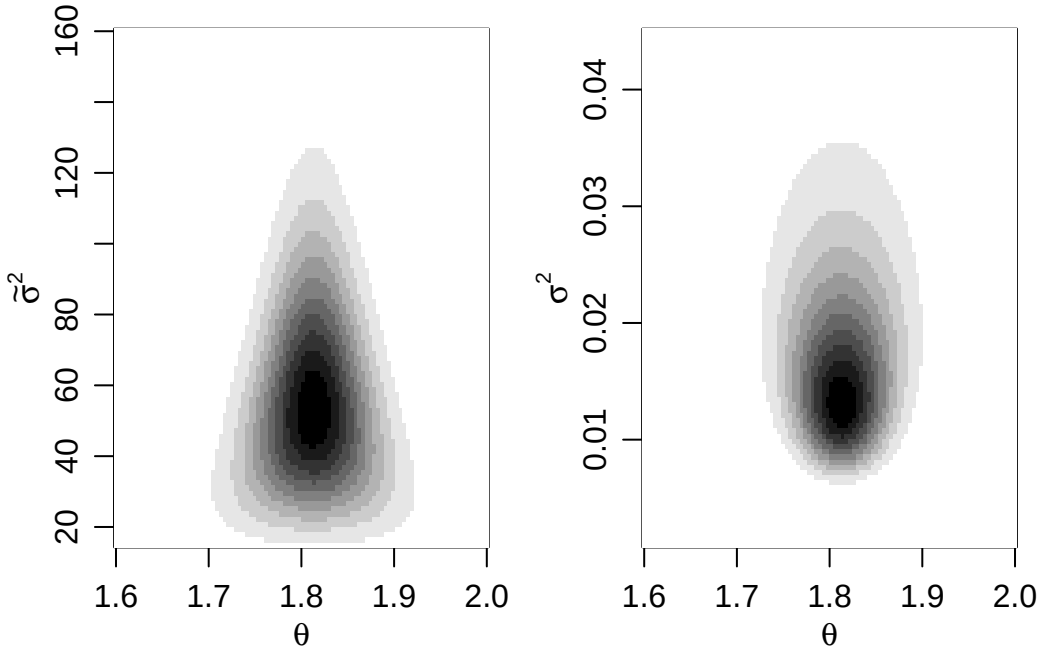
```
s2n
```

```
[1] 0.01702222
```

同時事後分布は $\mu_n = 1.814, \quad \kappa = 10, \quad \sigma_0^2 = 10$ の値によって、完全に定まり次のように表される。

$$\{\theta|y_1, \cdots, y_n, \sigma^2\} \sim normal(1.814, \sigma^2/10)$$
$$\{1/\sigma^2|y_1, \cdots, y_n\} \sim gamma(10/2, 10 \times 0.015/2)$$

ここで、$\tilde{\sigma}^2 = 1/\sigma^2$ とおくと、下図左のようなプロットが構築される。



4.4 多変量正規モデル

(後日加筆)

5 近似論 (理論篇)

5.1 モンテカルロ法

θ を興味あるパラメータとし、$y_1, \cdots, y_n$ を分布 $p(\theta|y_1, \cdots, y_n)$ からの標本の実現値とする。事後分布から独立に S 個のランダムな θ の値が生成できたとする。

$$\theta^{(1)}, \cdots, \theta^{(S)} \overset{\text{iid}}{\sim} p(\theta|y_1, \cdots, y_n)$$

このとき、$\{\theta^{(1)}, \cdots, \theta^{(S)}\}$ の経験分布は $p(\theta|y_1, \cdots, y_n)$ を近似しており、S を大きくすると近似は良くなることが示されている。$\{\theta^{(1)}, \cdots, \theta^{(S)}\}$ の経験分布は $p(\theta|y_1, \cdots, y_n)$ の**モンテカルロ近似**と呼ばれる。

$g(\theta)$ を θ の任意の関数とし、$\theta^{(1)}, \cdots, \theta^{(S)}$ を $p(\theta|y_1, \cdots, y_n)$ からの独立同一標本とすると、大数の法則により次が成り立つ。

$$\frac{1}{S}\sum_{i=1}^{S} g(\theta^{(s)}) \to E[g(\theta)|y_1, \cdots, y_n] = \int g(\theta)p(\theta|y_1, \cdots, y_n)d\theta \quad (S \to \infty)$$

これより、$S \to \infty$ の時、次のこともわかる。

* $\bar{\theta} = \sum_{i=1}^{S} \theta^{(s)}/S \to E[\theta|y_1, \cdots, y_n]$,

* $\sum_{i=1}^{S}(\theta^{(s)}-\bar{\theta})^2/(S-1) \to Var[\theta|y_1,\cdots,y_n]$,

* $\#\ (\theta^{(s)} \leq c)/S \to Pr(\theta \leq c|y_1,\cdots,y_n)$,

* 経験分布 $\to p(\theta|y_1,\cdots,y_n)$,

* メディアン $\to \theta_{1/2}$,

* α 分位点 $\to \theta_\alpha$

事後分布について関心のある統計量の多くは、十分な大きさのモンテカルロ標本を用いてほぼ正確に近似できる。

5.2 ギブスサンプラー

パラメータが複数あるモデルでは多くの場合、同時事後密度は標準的な分布にはならず、そこから直接サンプリングすることは難しい。それでも、各パラメータの完全条件付き分布からのサンプリングは容易であることが多い。そのような場合にはギブスサンプラーによる事後分布の近似が可能である。これは目標となる同時事後分布に収束するようなパラメータの値の従属列を構成する反復アルゴリズムである。本章では準共役な事前分布を用いた正規モデルの文脈でギブスサンプラーを概説し、この方法により事後分布がどれだけ上手く近似できるかを議論する。

5.2.1 準共役な事前分布

正規分布の章では θ の不確実性を σ^2 に依存する形で以下のようにモデル化した。

$$p(\theta|\sigma^2) = dnorm(\theta, \mu_0, \sigma/\sqrt{\kappa}_0)$$

この事前分布では θ の分散がデータの標本分散と関係しており、μ_0 は母集団からの κ_0 個の事前のサンプルの標本平均だと思うことができる。このモデルが妥当である状況もあるが、一方で θ の不確実性を σ^2 とは独立に特定化したい, つまり $p(\theta,\sigma^2) = p(\theta) \times p(\sigma^2)$ としたい場合もある。そのような同時分布の一つは以下の**準共役**な事前分布である。

$$\theta \sim gamma(\mu_0, \tau_0^2)$$
$$1/\sigma_n^2 \sim gamma(\nu_0/2, \nu_0\sigma_0^2/2)$$

ここで、

$$\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}, \quad \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)$$

である。

もし、τ_0^2 が σ^2 に比例しているならば、この事前分布は共役となり、$p(\sigma^2|y_1,\cdots,y_n)$ は逆ガンマ分布の密度になり、$\{\theta,\sigma^2\}$ の同時事後分布からのモンテカルロ標本を以下のように得ることができる。 1. $\sigma^{2(s)}$ の値を逆ガンマ分布 $p(\sigma^2|y_1,\cdots,y_n)$ から得る。

2. $\theta^{(s)}$ の値を正規分布 $p(\theta|\sigma^{2(s)},y_1,\cdots,y_n)$ から得る。

しかし、τ_0^2 が σ^2 に比例していない場合には、$1/\sigma^2$ の周辺分布はガンマ分布にならない。それどころかサンプリングが容易であるような標準的な分布にならない。

5.2.2 ギブスサンプラー

分布 $p(\theta|\sigma^2, y_1, \cdots, y_n)$ 及び $p(\sigma^2|y_1, \cdots, y_n)$ は、それぞれ θ と σ^2 以外の全ての変数を条件づけていることから**完全条件付き分布**と呼ばれる。反復サンプリングの考え方に基づき、パラメータの現在の状態 $\phi^{(s)} = \{\theta^{(s)}, \tilde{\sigma}^{2(s)}\}$ が与えられている時、新たな状態を次のようにして生成する。

1. $\theta^{(s+1)} \sim p(\theta|y_1, \cdots, y_n)$ を生成する。
2. $\tilde{\sigma}^{2(s+1)} \sim p(\tilde{\sigma}^2|\theta^{(s+1)}, y_1, \cdots, y_n)$ を生成する。
3. $\phi^{(s+1)} = \{\theta^{(s+1)}, \tilde{\sigma}^{2(s+1)}\}$ とする。

このアルゴリズムは**ギブスサンプラー**と呼ばれ、パラメータ $\{\phi^{(1)}, \cdots, \phi^{(S)}\}$ の独立でない列を発生させる。

先ほどの例では、正規モデルを題材にギブスサンプラーを概観したが、ここからはより一般的な性質について述べていく。今パラメータのベクトル $\phi = \{\phi_1, \cdots, \phi_p\}$ があるとして、ϕ に関する情報が $p(\phi) = p(\phi_1, \cdots, \phi_p)$ で測られているとしよう。初期値 $\phi^{(0)} = \{\phi_1^{(0)}, \cdots, \phi_p^{(0)}\}$ を所与として、以下のギブスサンプラーにより、$\phi^{(s)}$ を $\phi^{(s-1)}$ から生成する。

1. $\phi_1^{(s)} \sim p(\phi_1|\phi_2^{(s-1)}, \phi_3^{(s-1)}, \cdots, \phi_p^{(s-1)}$ を生成する。
2. $\phi_1^{(s)} \sim p(\phi_2|\phi_1^{(s-1)}, \phi_3^{(s-1)}, \cdots, \phi_p^{(s-1)}$ を生成する。

⋮

p. $\phi_p^{(s)} \sim p(\phi_1|\phi_2^{(s)}, \phi_3^{(s)}, \cdots, \phi_{p-1}^{(s)}$ を生成する。

このアルゴリズムは以下の「独立でない」ベクトルの列を生成する。

$$\begin{aligned}
\phi^{(1)} &= \{\phi_1^{(1)}, \cdots, \phi_p^{(1)}\}, \\
\phi^{(2)} &= \{\phi_1^{(2)}, \cdots, \phi_p^{(2)}\}, \\
&\vdots \\
\phi^{(S)} &= \{\phi_1^{(S)}, \cdots, \phi_p^{(S)}\}
\end{aligned}$$

この列において $\phi^{(s)}$ は $\phi^{(s-1)}$ のみを通じて $\phi^{(0)}, \cdots, \phi^{(s-1)}$ に依存している。すなわち、$\phi^{(s-1)}$ を所与として、$\phi^{(s)}$ は $\phi^{(0)}, \cdots, \phi^{(s-2)}$ から条件付き独立である。この性質はマルコフ性と呼ばれ、この列は**マルコフ連鎖**と呼ばれる。ある条件のもとで、以下が成立する。

$$Pr(\phi^{(s)} \in A) \to \int_A p(\phi) d\phi \quad (s \to \infty)$$

言葉で説明すると、$s \to \infty$ の時、初期値 $\phi^{(0)}$ に関わりなく、標本分布は目標分布に収束するということになる。さらに、重要なこととして、関心のあるほとんどの関数 g に対して、以下が成り立つ。

$$\frac{1}{S}\sum_{s=1}^{S} g(\phi^{(s)}) \to E[g(\phi)] = \int g(\phi)p(\phi)d\phi \quad (S \to \infty)$$

これはモンテカルロ近似のように、$\{g(\phi^{(1)}), \cdots, g(\phi^{(S)})\}$ の標本平均で $E[g(\phi)]$ を近似できることを意味している。そのため、この近似は**マルコフ連鎖モンテカルロ**近似と呼ばれ、計算は MCMC アルゴリズムと呼ばれる。

モンテカルロ法を用いたベイズ的データ分析では、サンプリングの手順と確率分布とが複雑にならび混乱が生じることが多い。このことを念頭に、データ分析のうち統計学的な部分と、数値近似に関する部分とを峻別することは理解の助けになる。ベイズ的データ分析に必要な構成要素とは

1.「モデルの特定」　確率分布の集合の特定、ある値のもとでデータの標本分布を表現する。
2.「事前分布の特定」　どのようなパラメータの値が標本分布をよく記述するかに関する、誰かの事前の情報を表す。
3.「事後分布の要約」　関心のある特定の値をによってなされる (代表値、信用区間等)

多くのモデルでは、$p(\phi|y)$ は複雑になり、書き下すのも大変である。このような場合に事後分布を見てみるには、$p(\phi|y)$ からのモンテカルロ標本を調べるのが有用である。したがって、モンテカルロ及び MCMC サンプリングのアルゴリズムとは、

- モデルではなく
- $\mathbf{y}$ と $p(\theta)$ がもつ情報以上のものを生み出すことはなく、
- $p(\phi|y)$ を見るための単なる手段に過ぎない

5.3　メトロポリスアルゴリズム

(省略)

6　階層モデリング

6.1　階層データ

この章で使うデータは、米国の二つの公立高校の一年生の数学の得点を表している。学校 1 から 31 人、学校 2 から 28 人の生徒が無作為に選ばれ、数学試験に参加している。両校とも 1 年生の生徒数はそれぞれ約 600 人で、どちらも都市部に位置している。学校 1 の一年生全員がテストを受けた場合に得られる平均点 θ_1 の推定や、それを学校 2 での平均点 θ_2 と比較することに興味があるとする。

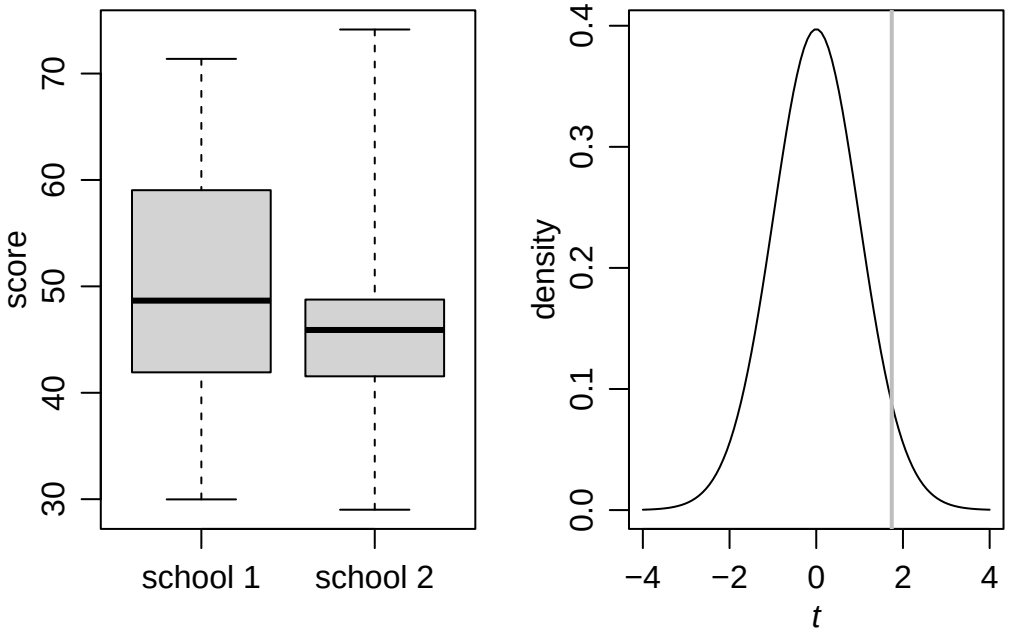
```
[1] 50.81355
```

```
[1] 46.15071
```

```
[1] 10.44635
```

```
    Welch Two Sample t-test

data:  y1 and y2
t = 1.7612, df = 56.288, p-value = 0.08363
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.640171  9.965839
sample estimates:
mean of x mean of y
 50.81355  46.15071
```



このような入れ子となった集団の構造を持つデータを**階層データ**や**マルチレベルデータ**と呼ばれる。同じようなデータ構造を持つ例としては以下のようなものが挙げられる。

- 複数の病院に入院している患者のデータ.
- 動物のあるグループ内の遺伝子
- ある国の地域の郡に属する人々

最も単純な階層データは二つのレベルをもち、一つ目のレベルは「グループ」で構成されており、二つ目のレベルが「グループ内ユニット」となるものである。この場合、グループ j 内の i 番目のユニットのデータを $y_{i,j}$ のように表す。

階層データ $\{Y_1, \cdots, Y_m\}$, $\quad Y_j = \{Y_{1,j} \cdots, Y_{n_j,j}\}$ を記述するためのモデルについて考えてみよう。モデル $p(y_1, \cdots, y_m)$ はどのような性質を満たすべきだろうか。まず、あるグループ j からのデータ周辺確率密度である $p(y_j) = p(y_{1,j}, \cdots, y_{n_j,j})$ について考えてみる。グループ数 j に相当する母集団が、実際のサンプルサイズ

n_j に比べて大きい場合、グループ固有のパラメータ ϕ_j を用いてグループ j 内のデータを以下のように条件付き独立同一な形でモデル化することができる。

$$\{Y_{1,j}, \cdots, Y_{n_j,j} | \phi_j\} \overset{\text{iid}}{\sim} p(y|\phi_j)$$

ここで、$\phi_1, \cdots, \phi_m$ に関する情報をどのように表現すれば良いだろうか。前述のように、これらのパラメータを独立したものとして扱うのは妥当でないように思える。なぜなら独立性のもとでは $\phi_1, \cdots, \phi_{m-1}$ の値を知っていても ϕ_m に関する情報は変わらないからである。しかし、グループ自体がより大きなの母集団からのサンプルである場合、グループ固有のパラメータの交換可能性は適切であろう。デ・フィネッティの定理を適用とすると、未知のパラメータ ψ に依存したモデル $p(\phi|\psi)$ を用いて

$$\{\phi_1, \cdots, \phi_m | \psi\} \overset{\text{iid}}{\sim} p(\theta|\psi)$$

と表現できる。

このデ・フィネッティの定理の二重適用によって以下の三つの確率分布を導出することができる。

$$\begin{aligned} \{y_{1,j}, \cdots, y_{n_j,j} | \phi_j\} &\overset{\text{iid}}{\sim} p(y|\phi_j) \quad (change\ in\ group) \\ \{\phi_1, \cdots, \phi_m | \psi\} &\overset{\text{iid}}{\sim} p(\phi|\psi) \quad (change\ among\ groups) \\ \psi &\sim p(\psi) \qquad (prior) \end{aligned}$$

上のモデル式 1,2 段目は標本分布と呼ばれ、3 段目は事前分布として区別される。実際に、前者はデータを用いて推定される一方で、事前分布はデータから推定されないことに注意する。

6.1.1 階層正規モデル

複数の集団における平均の不均一性を記述するための有名なモデルとして階層正規モデルがある。これはグループ内とグループ間の確率モデルが以下のような正規分布で表現されるモデルである。

$$\begin{aligned} \phi_j = \{\theta_j, \sigma^2\}, \quad p(y|\phi_j) = normal(\theta_j, \sigma^2) \quad (in\ group\ model) \\ \psi = \{\mu, \tau^2\}, \quad p(\theta_j|\psi) = normal(\mu, \tau^2) \quad (among\ group\ model) \end{aligned}$$

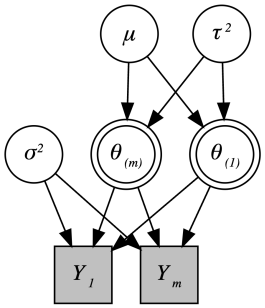


Figure1: 階層正規モデル 1

ここでは、グループ内の変動 σ^2 はグループ同一で一定であると仮定するが、不均一な分散構造を表現するモデルも扱うことができる (後述) このモデルの未知のパラメータは μ, τ^2, σ^2 である。便宜上、以下のような準共役の正規分布と逆ガンマ事前分布を用いる。

$$
\begin{aligned}
1/\sigma^2 &\sim gamma(\nu_0/2, \nu_0\sigma_0^2/2) \\
1/\tau^2 &\sim gamma(\eta_0/2, \eta_0\tau_0^2/2) \\
\mu &\sim normal(\mu_0, \gamma_0^2)
\end{aligned}
$$

6.2 事後推測

階層正規モデルがもつ未知の量は、グループ固有の平均値 $\{\theta_1, \cdots, \theta_m\}$ グループ内の変動 σ^2、グループ固有の平均値の母集団の平均と分散 (μ, τ^2) である。これらのパラメータの同時事後推論を行うため、ギブスサンプラーによって、事後分布 $p(\theta_1, \cdots .\theta_m, \mu, \tau^2\sigma^2|y_1, \cdots, y_m)$ を近似する。

一変量正規モデルとの類似性を認識することが重要であるので、事後分布を以下のように分解する。

$$
\begin{aligned}
&p(\theta_1, \cdots .\theta_m, \mu, \tau^2\sigma^2|y_1, \cdots, y_m) \\
&\propto p(\mu, \tau^2\sigma^2) \times p(\theta_1, \cdots .\theta_m|\mu, \tau^2\sigma^2) \times p(y_1, \cdots, y_m|\theta_1, \cdots .\theta_m, \mu, \tau^2\sigma^2) \\
&= p(\mu)p(\tau^2)p(\sigma^2) \left\{ \prod_{j=1}^{m} p(\theta_j|\mu|\tau^2) \right\} \left\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma^2) \right\}
\end{aligned}
$$

２番目のカッコ内の項は、階層正規モデルの条件付き分布独立性による結果である。$\{\theta_1, \cdots .\theta_m, \mu, \tau^2\sigma^2\}$ が所与のもと、確率変数 $Y_{1,j}, \cdots, Y_{n_j,j}$ は互いに独立で θ_j と σ^2 のみに依存し、μ や τ^2 二は依存しない分布に従う。このような構造はグラフをみると直感的に読み取ることができる。(μ, τ^2) から各 Y_j へのパスは、(μ, τ^2) が Y_j に影響を与えていることを意味するが、それは θ_j を介した間接的なものであり、θ_j によって両者は分断されていることが読み取れる。各パラメータの完全条件付き分布は以下に示す通りである。

$$
\begin{aligned}
\{\mu|\theta_1, \cdots, \theta_m, \tau^2\} &\sim normal\left(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1\gamma_0^2}, [m/\gamma^2 + 1/\gamma_0^2]^{-1} \right) \\
\{1/\tau^2|\theta_1, \cdots, \theta_m, \mu\} &\sim gamma\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum(\theta_j - \mu)^2}{2} \right) \\
\{\theta_j|y_{1,j}, \cdots, y_{n_j,j}, \mu, \tau^2, \sigma^2\} &\sim normal\left(\frac{n_j\bar{y}_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, [n_j/\sigma^2 + \mu/\tau_0^2]^{-1} \right) \\
\{1/\sigma^2|\theta, y_1, \cdots, y_m\} &\sim gamma\left(\frac{1}{2}\left[\nu_0 + \sum_{j=1}^{m} n_j \right], \frac{1}{2}\left[\nu_0\sigma_0^2 + \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{i,j} - \theta_j)^2 \right] \right)
\end{aligned}
$$

不均一分散の階層モデルを図示すると以下のようになる。

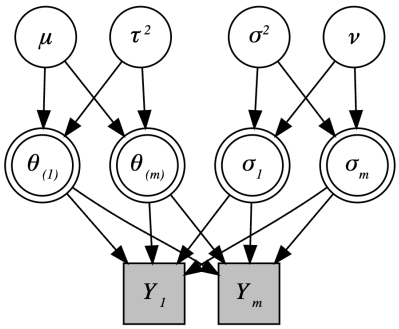


Figure2: 不均一分散を持つ階層正規モデル

7　ベイズ回帰モデル (多変量執筆後)

7.1　線形回帰

7.2　階層回帰