Rで学ぶベイズ統計学入門

3.4　ベイズ法の頑健性について
3.5　共役事前分布の混合形
3.6　コインの偏りについてのベイズ検定

茨城大学工学部
佐々木研究室
荒井悠有

３．４ベイズ法の頑健性について

ー事前情報に適合する異なる事前分布が複数存在するとき、その選択によって事後分布が大きな影響を受けない場合、ベイズ分析は事前分布の選択に対して頑健であるといわれる.

・具体例

ジョーという人物のIQ（＝θ）の真の値を推定する.

事前分布として正規分布とt分布を考える.

ジョーは４回IQテストを受け、そのスコアがy1,y2,y3,y4とする.

それぞれのスコアyの分布が、既知の標準偏差σ＝15の正規分布N(θ,σ)に従うと仮定すると、観測されるスコアの平均値 $\bar{y}$ はN(θ,σ/$\sqrt{4}$)となる.

-事前情報（事前の確信）

・ジョーのIQは平均的で、その事前分布の中央値は100

・90%の確率でジョーのIQは80と120の間に位置する

・IQの中央値が100で95％分位点が120である

事前情報に適合する正規密度の平均値と標準偏差値を求める

→normal.select()関数を使用

正規分布の二つの分位点について情報があるとき、これに適合する正規密度の平均と標準偏差を見つける

```
quantile1 <- list(p = .5, x= 100);
quantile2 <- list(p = .95, x = 120)
normal.select(quantile1, quantile2)

$mu
[1] 100

$sigma
[1] 12.15914
```

平均μ＝100で標準偏差τ＝12.16の正規密度が、事前情報に適合する

例-1：正規分布(正規事前分布)

観測されるテストスコアの平均を仮に $\overline{y}$ ＝110、$\overline{y}$ ＝125、$\overline{y}$ ＝140とし、それぞれの場合についてジョーの真のIQ(=θ)の事後平均mu1と事後標準偏差tau1を計算する.

事後の精度P_1 ＝$4/\sigma^2$＋$1/\tau^2$

事後の標準偏差τ_1 ＝1/ $\sqrt{P_1}$ ＝1/($\sqrt{4/\sigma^2+1/\tau^2}$)

Θの事後平均μ_1 ＝ $$\frac{\overline{y}(4/\sigma^2)+\mu(1/\tau^2)}{4/\sigma^2+1/\tau^2}$$

	ybar	mu1	tau1
[1,]	110	107.2442	6.383469
[2,]	125	118.1105	6.383469
[3,]	140	128.9768	6.383469

例-2:t分布(t事前分布)

・利用するt密度

-位置パラメータμ （中央値μ=100)
-尺度パラメータτ
-自由度２

ここで、τはθの95%分位点が120であるように設定する.

$$P(\theta < 120) = P(T < 20/\tau) = 0.95$$

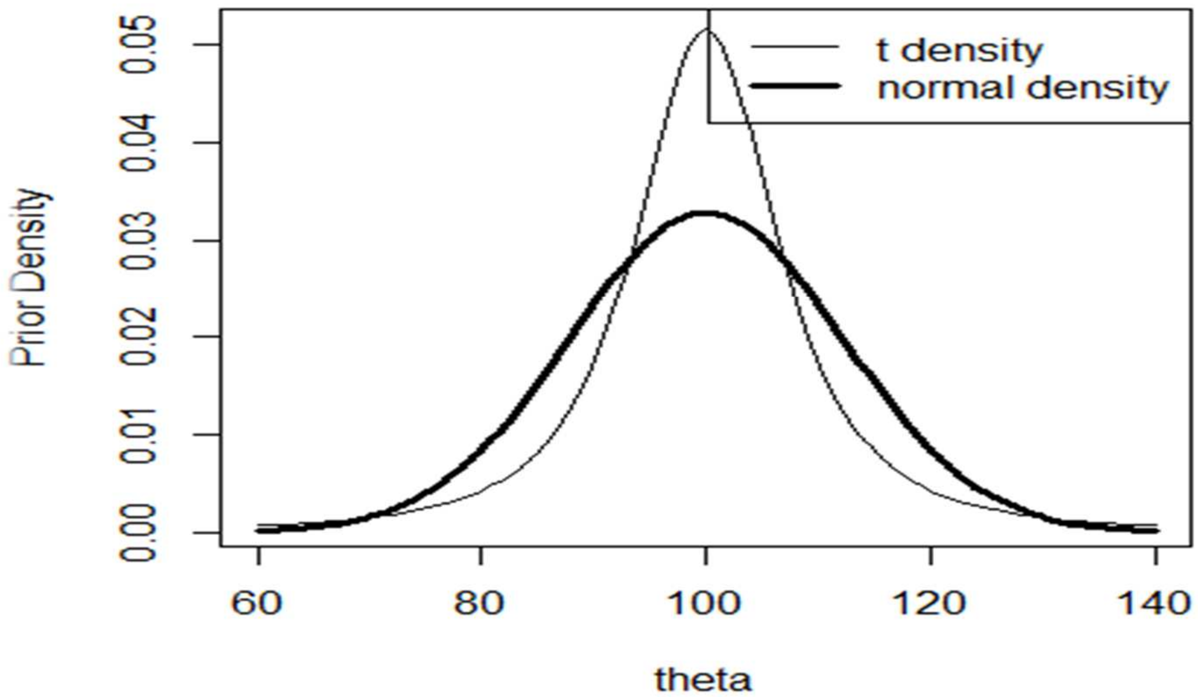
Tは自由度2の標準t変量である。したがって次の式が成り立つ.

$$\tau = 20/t_2(0.95)$$

ここで$t_v(p)$は自由度vのtランダム変数におけるp分位点の値である.
t分位点関数qt()を使ってτを以下のように求める.

```
>tscale <- 20/qt(0.95,2)
>tscale
[1] 6.849349
```

正規事前密度とt事前密度をグラフに表示すると以下のようになる。



テストスコアが極端な値の場合、事後密度に影響がある.

t事前密度を使って事後密度を計算する.
θの事後密度は比例定数を除いて以下で与えられる.

$$g(\theta \mid data) \propto \phi(\overline{y} | \theta, \sigma / \sqrt{n}) gT(\theta \mid v, \mu, \tau)$$

ϕ(y|θ,σ) : 平均θで標準偏差σの正規密度

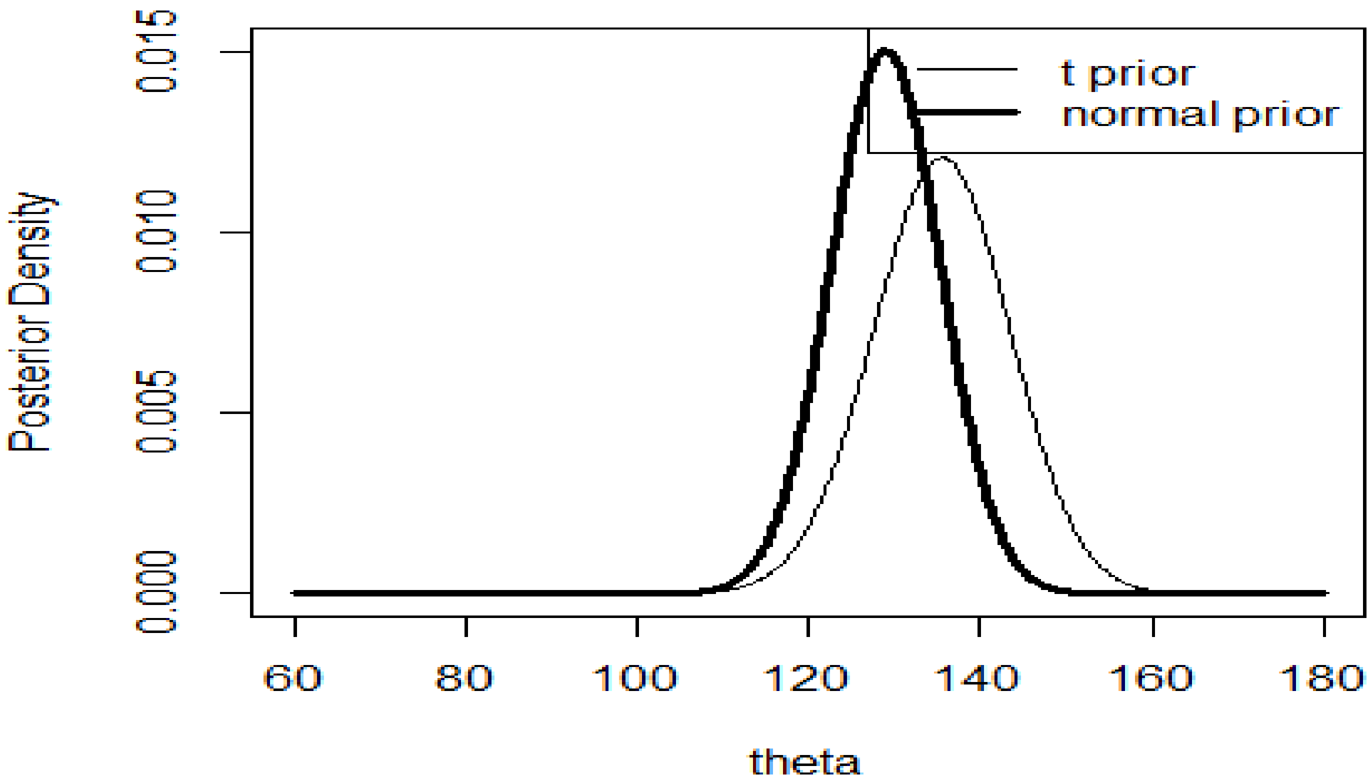
gT(μ|v,μ,τ):中央値μで尺度パラメータτ、自由度vのt密度

	ybar	mu1 t	tau1 t
[1,]	110	105.2921	5.841676
[2,]	125	118.0841	7.885174
[3,]	140	135.4134	7.973498

二つの事前密度(正規事前密度とt事前密度)を利用したΘの事後積率を比較.

	ybar	mu1	tau1	ybar	mu1 t	tau1 t
[1,]	110	107.2442	6.383469	110	105.2921	5.841676
[2,]	125	118.1105	6.383469	125	118.0841	7.885174
[3,]	140	128.9768	6.383469	140	135.4134	7.973498

観測されるスコアの平均が事前の確信と近い場合、事後平均値と事後標準偏差値は、どちらの事前密度を使っても近い値になっている.
しかし、極端な値をとる場合、事前分布のとり方によって違いがみられる.



スコア平均が極端な値をとる場合における二つの事前密度に対する事後密度のグラフ

･まとめ

-観測される平均スコアが事前の確信と矛盾しない場合
→推論は、事前分布の選択に対して頑健である

-矛盾する場合(極端な値をとる場合)
→推論は事前密度の選択に対して頑健ではない

3.5 共役事前分布の混合形

共役事前分布：事前分布と事後分布が同じ関数形を持つ.

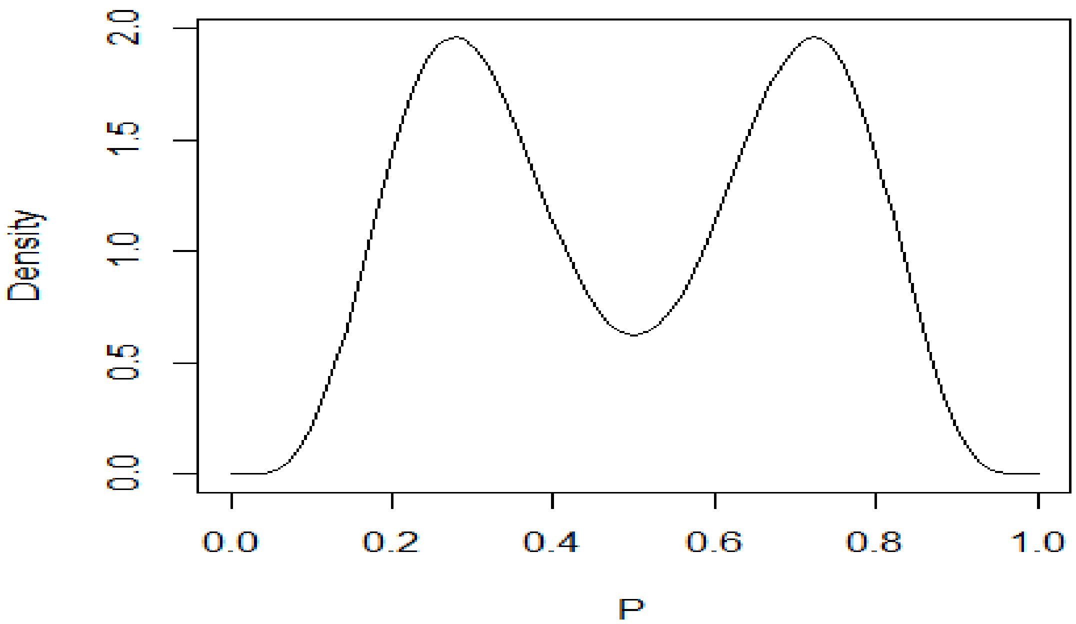
ここでは、この共役事前分布を拡張する方法の一つとして
離散混合形を使うことを説明する.

例として、ベータ密度の混合形を使って偏ったコインで表が出る
確率を調べる方法を紹介する.

コインの表が出る確率をpとする.
pは0.3か0.7の付近にあり、そのどちらかに位置する確率は等しい.

$$g(p)=\gamma g_1(p)+(1-\gamma)g_2(p)$$

g_1:ベータ分布beta(6,14)
g_2 : ベータ分布beta(14,6)
混合確率$\nu=0.5$



コインをn回投げて、表がs回、裏がf=n-s回でるとする.
この割合の事後密度は次の混合形となる.

$$g(p|\mathrm{data}) = \gamma(\mathrm{data})g_1(p|\mathrm{data}) + (1-\gamma(\mathrm{data}))g_2(p|\mathrm{data})$$

g_1: ベータ分布beta(6+s,14+f)
g_2: ベータ分布beta(14+s,6+f)

混合確率γ(data)は次の形になる.

$$\gamma(\mathrm{data}) = \frac{\gamma f_1(s,f)}{\gamma f_1(s,f) + (1-\gamma) f_2(s,f)}$$

$f_i(s,f)$: pの事前密度がg_jの場合にコインをn回投げて表がs回出る事前予測確率

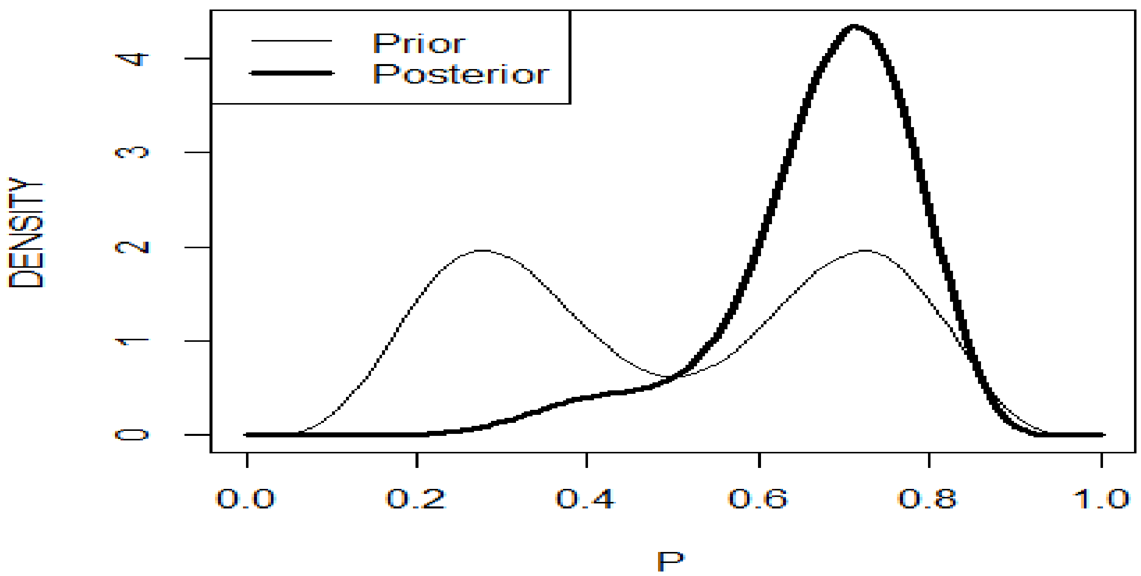
binomial.beta.mix()関数を用いて、γ(data)の値を求める.

```
> probs <- c(.5, .5)
> beta.par1 <- c(6,14)
> beta.par2 <- c(14,6)
> betapar <- rbind(beta.par1, beta.par2)
> data <- c(7,3)
> post <- binomial.beta.mix(probs,betapar,data)
> post
$probs
 beta.par1   beta.par2
0.09269663  0.90730337

$betapar
          [,1] [,2]
beta.par1   13   17
beta.par2   21    9
```

10回コインを投げ、表が7回、裏が3回出るとする.
先ほどの出力から、pの事後分布は以下のベータ分布の混合形になる.

$$g(p|data) = 0.093beta(13,17) + 0.907beta(21,9)$$



偏ったコインの場合の割合についての事前密度と事後密度

3.6 コインの偏りについてのベイズ検定

yがパラメータnとpの二項分布にしたがうことを観測する.
コインの表が出る確率p=0.5とする仮説Hを検定する.
yが観測されるのならば、通常は次のp値に基づいて判断する.

$$2\times\min\{P(Y\le y), P(Y\ge y)\}$$

このp値が小さい場合、仮説Hを棄却し、コインには偏りがあると結論する.

例えば・・・
コインを20回投げるとして表が出るのは5回のみとする.
この時の確率は

```
>pbinom(5, 20, 0.5)
[1] 0.02069473
```

この場合p値は2×0.021=0.042となり、通常の有意水準0.05よりも小さいので、仮説Hを棄却し、コインには偏りがあると結論する.

ベイズの観点から考えると次の二つのモデルがある.

・コインには偏りがない(p=0.5)
・コインには偏りがある(p≠0.5)

よって、この検定において事前分布は次の混合形で表せる.

$$g(p) = 0.5I(p=0.5) + 0.5I(p\neq0.5)g_1(p)$$

g1(p): (0,1)上にあるp=0.5とは異なる事前分布
I(A): 指示関数.事象Aが真ならば1、偽ならば0

n回の試行で表の出る回数を確認し、ベイズの法則に基づいて最初の事前分布を更新する.pの事後密度は次のように表せる.

$$g(p|y) = \lambda(y)I(p=0.5) + (1-\lambda(y))g_1(p|y)$$

g_1はベータ密度beta(a+y,a+n-y)であり、λ(y)はコインに偏りがないとするモデルの事後確率で以下のように表せる.

$$\lambda(y) = \frac{0.5p(y|0.5)}{0.5p(y|0.5)+0.5m_1(y)}$$

p(y|0.5): p=0.5の場合のyの二項密度
$m_1(y)$: ベータ密度を利用したyについての（事前）予測密度

$$m_1(y) = \frac{f(y|p)g_1(p)}{g_1(p|y)}$$

コインに偏りがある場合のpにベータ事前分布beta(10,10)を割り当て、n=20回投げてy=5回の表を観測する.

```
> n <- 20
> y <- 5
> a <- 10
> p <- 0.5
> m1 <- dbinom(y,n,p) * dbeta(p,a,a)/dbeta(p,a+y,
+ a+n-y)
> lambda <- dbinom(y,n,p)/(dbinom(y,n,p)+m1)
> lambda
[1] 0.2802215
```

LearnBayesパッケージのpbetat()関数は二項割合の検定を行う.
引数として検定したいpの値、その値が正しいと信じる事前確率、その値が正しくないときpがどのように分布するかに関するβ分布のパラメータ、表の回数と裏の回数.

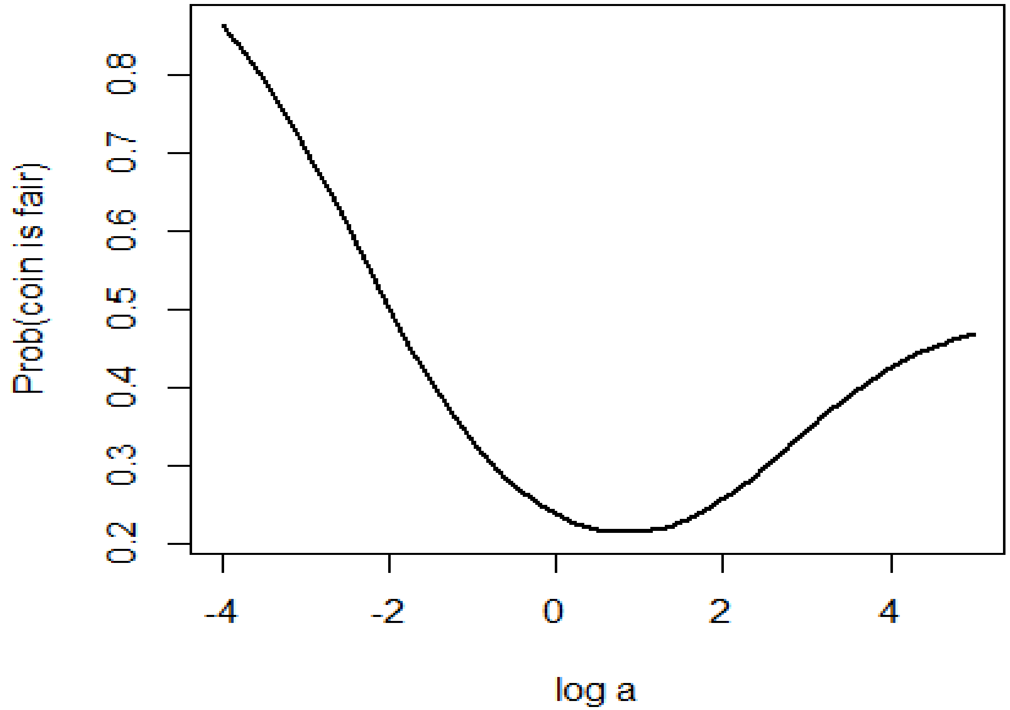
```
>pbetat(p, .5, c(a,a), c(y, n-y))

$bf
[1] 0.3893163

$post
[1] 0.2802215
```

postはp=0.5の事後確率
bfはベイズファクターの値（詳しくは8章で）

事前分布のパラメータaの選択によって事後分布の計算が影響されるのか？



どのような事前分布に対しても、事後確率は0.2よりも大きくなる.

→先ほどのp値(0.042)はコインに偏りはないとする仮説を棄却する証拠として過大な値である.

頻度主義とベイズ法の計算の違い

頻度主義:「表が5回以下」である事象の確率に基づく
ベイズ法:「表がちょうど5回」である事象の尤度だけに基づく

疑問:「表が5回以下」だとすると、ベイズ法による答えに変化はあるのか?
この場合、コインに偏りがない事後確率は以下で求められる.

$$\lambda(y) = \frac{0.5P_0(Y \le 5)}{0.5P_0(Y \le 5) + 0.5P_1(Y \le 5)}$$

$P_0(Y \le 5)$:p=0.5の二項モデルのもとで表が5回以下現れる確率
$P_1(Y \le 5)$:ベータ事前分布beta(10,10)のもとでのこの事象の予測確率

二項モデルのもとで表が5回現れる累積確率を求める.

```
> n <- 20
> y <- 5
> a <- 10
> p <- .5
> m2 <- 0
> for (k in 0:y)
+ m2 <- m2 + dbinom(k,n,p) * dbeta(p,a,a)/dbeta(p,a+k,
+ a+n-k)
> lambda <- pbinom(y,n,p)/(pbinom(y,n,p) + m2)
> lambda
[1] 0.2184649
```

出力より、「表が5回以下」というデータに基づいた場合、偏りがない事後確率は0.218.

「表が5回以下」ということは、「表がちょうど5回」よりも、偏りがないという仮説を棄却する十分な証拠になる.

よって、この値は合理的な結果と言える.