Rで学ぶベイズ統計学入門

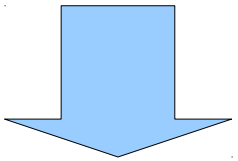
第2章
ベイズ的思考への誘い

新納 浩幸

本章の概要

母集団の割合を調査するという問題を通して、
ベイズ推論の基本的概念を解説

(問題)
アメリカの大学生の何割が8時間以上の
睡眠をとっているか?



母集団はアメリカの大学生全体 (全部調査は不可能)
推定対象は8時間以上睡眠をとっている学生の割合 p

ベイズ的思考(1)

27 人ランダムに調べて、 11人が8時間以上睡眠をとっていた

$$p = 11/27 = 0.4074$$

標準的な考え方、でも少し変

p の値は確率的に変化する

p は確率変数

ベイズ的な考え方

ベイズ的思考(2)

p は確率変数なので分布を持つ

$g(p)$ ← p の事前分布

p の推定は事後確率最大(MAP推定)

$$g(p|data) = \frac{g(p)g(data|p)}{g(data)}$$

上式を最大にする p が求める p

→ $g(p)g(data|p)$ を最大にする p

尤度

$g(data|p)$　　p の下で data が発生する確率

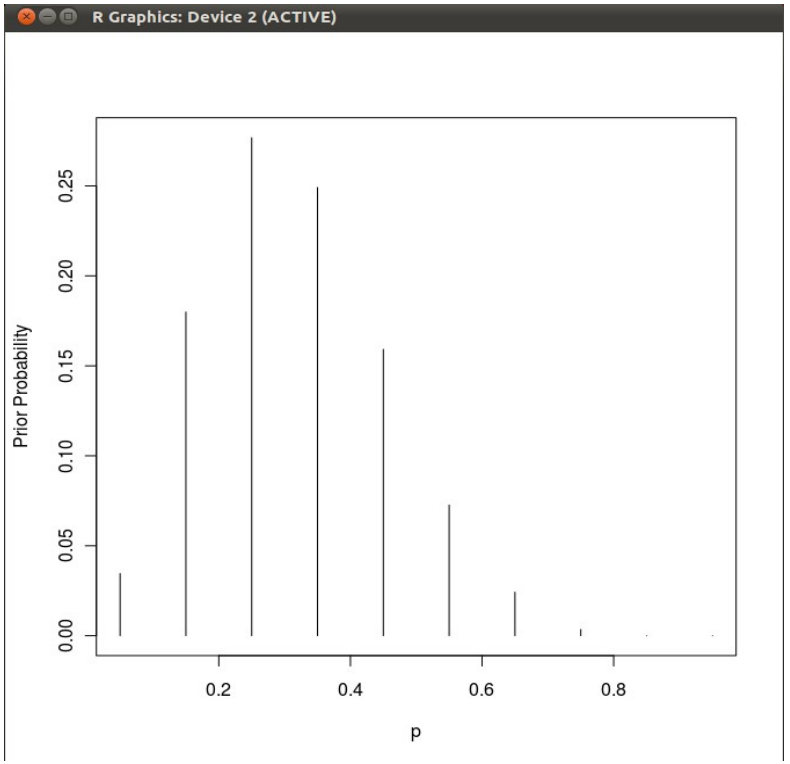
data： s 人が 8時間以上の睡眠、
f 人が 8時間未満の睡眠

$$g(data|p) =_{s+f} C_s p^s (1-p)^f$$

$$\propto p^s (1-p)^f$$

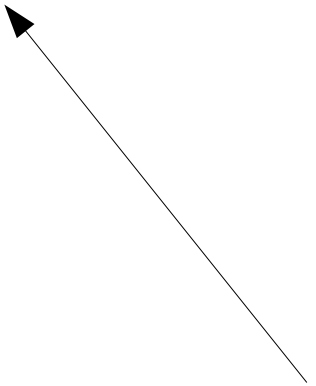
離散事前分布

```
> (prior <- c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0))
 [1] 1.0 5.2 8.0 7.2 4.6 2.1 0.7 0.1 0.0 0.0

> (prior <- prior/sum(prior))
 [1] 0.034602076 0.179930796 0.276816609 0.249134948 0.159169550 0.072664360
 [7] 0.024221453 0.003460208 0.000000000 0.000000000
```



事前確率と事後確率の比較

```
> post <- pdisc(p, prior, c(11,16))

> round(cbind(p, prior, post),2)
        p prior post
 [1,] 0.05  0.03 0.00
 [2,] 0.15  0.18 0.00
 [3,] 0.25  0.28 0.13
 [4,] 0.35  0.25 0.48
 [5,] 0.45  0.16 0.33
 [6,] 0.55  0.07 0.06
 [7,] 0.65  0.02 0.00
 [8,] 0.75  0.00 0.00
 [9,] 0.85  0.00 0.00
[10,] 0.95  0.00 0.00
```



先ほどの式での事後確率の計算
p = 0.05 ～ 0.95
prior　事前確率の分布
s = 11,　f = 16

連続型事前分布

$g(p)$ ⟵ 連続型事前分布

ベータ分布を利用する (共役事前分布となっているから)

$$g(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$$

$$B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1} dp$$ ベータ関数

a と b は未知パラメータ

a と b の決め方(1)

beta.select 関数の利用

2つの形状を表す情報を与えると
a と b が求まる

中央値 → 0.3

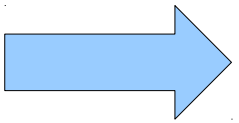
p はある範囲の値、だいたい 0.15 から 0.45 の間、中間は 0.3

90%分位点 → 0.5

p が 0.5 以下である確率は 0.9

a と b の決め方(2)

```
> q1 <- list(p = 0.9, x = 0.5)
> q2 <- list(p = 0.5, x = 0.3)
> beta.select(q1,q2)
[1] 3.26 7.18
```



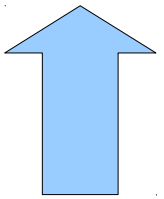
$$a = 3.26 \qquad b = 7.18$$

事後分布の形

$$g(p)g(data|p) = \frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}\ {}_{s+f}C_s p^s (1-p)^f$$

$$\propto p^{a+s-1}(1-p)^{b+f-1}$$

$$\propto \frac{1}{B(a+s,b+f)}p^{a+s-1}(1-p)^{b+f-1}$$



パラメータが a+s と b+f のベータ分布

p の解析

p が 0.5 以下になる確率

```
> pbeta(0.5,a+s,b+f)
[1] 0.9307817
```

```
> a <- 3.26
> b <- 7.18
> s <- 11
> f <- 16
```

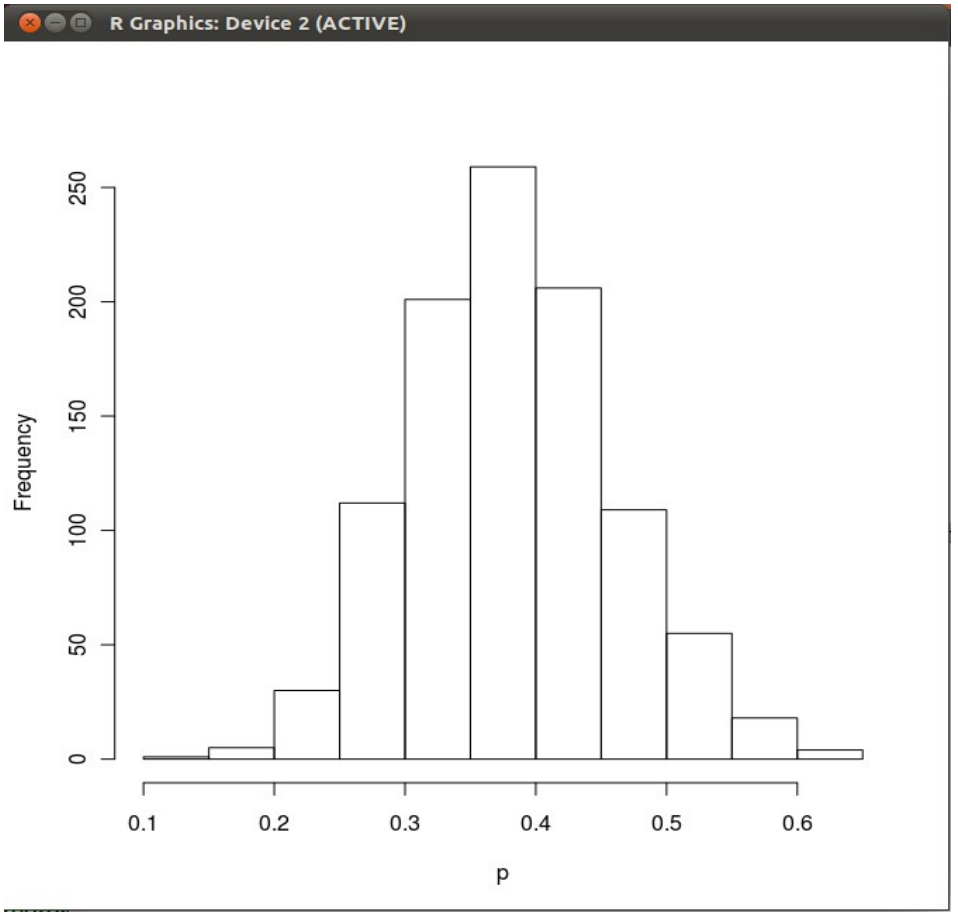
p の 90% 信頼区間

```
> qbeta(c(0.05,0.95),a+s,b+f)
[1] 0.2556027 0.5134840
```

シミュレーション

事後分布の計算が面倒な(できない)ときは、
事後分布に従った乱数を発生してシミュレーション

```
> ps <-rbeta(1000,a+s,b+f)
> hist(ps,xlab="p",main="")
```

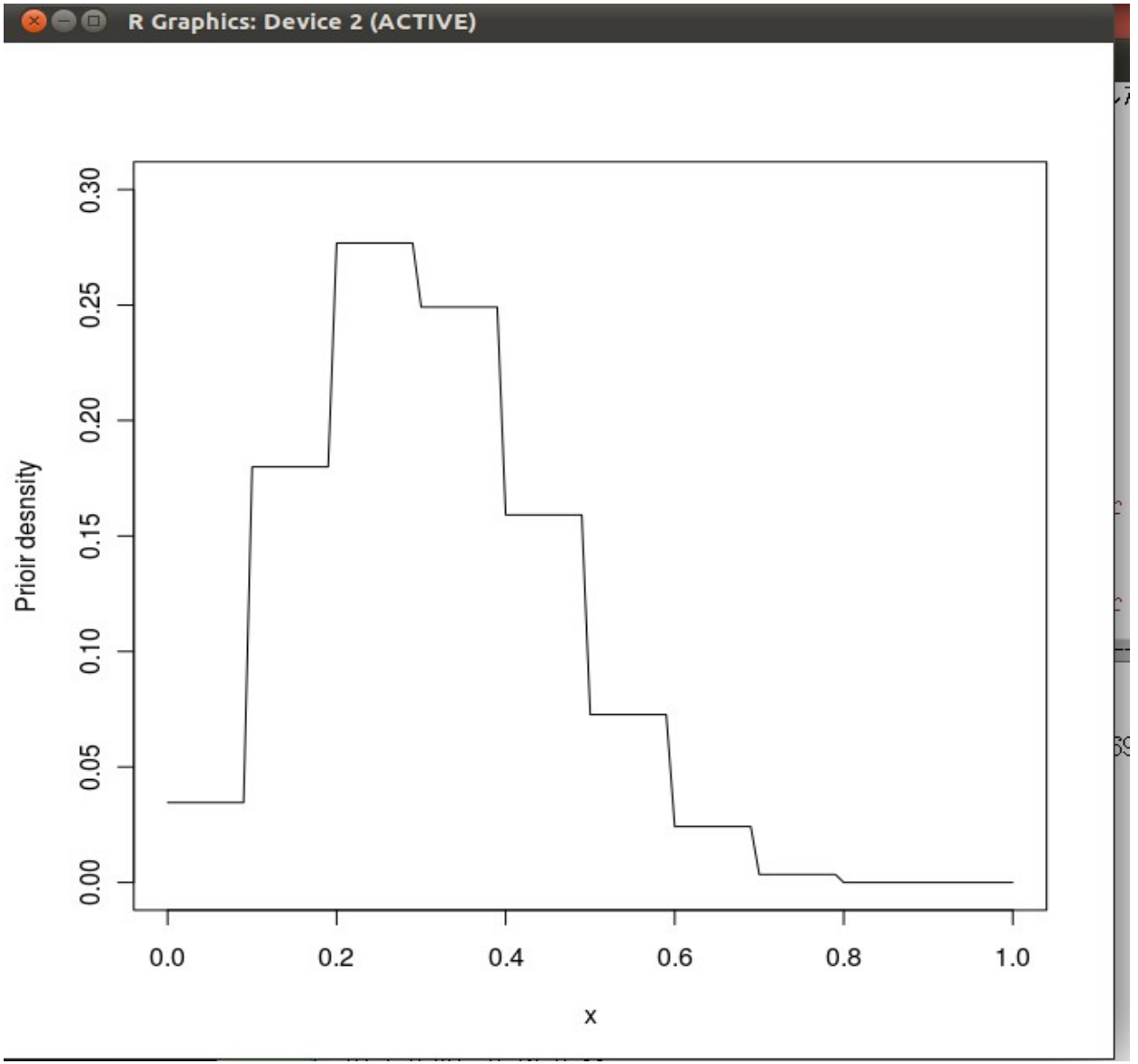


ヒストグラム事前分布(1)

離散型事前分布をヒストグラムにして、
それを連続型事前分布として扱う

```
> prior <- c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)
> prior <- prior/sum(prior)
> midpt <- seq(0.05, 0.95, by=0.1)
> curve(histprior(x,midpt,prior), from=0, to=1, ylab="Prioir desnsity", ylim = c(0,0.3))
```

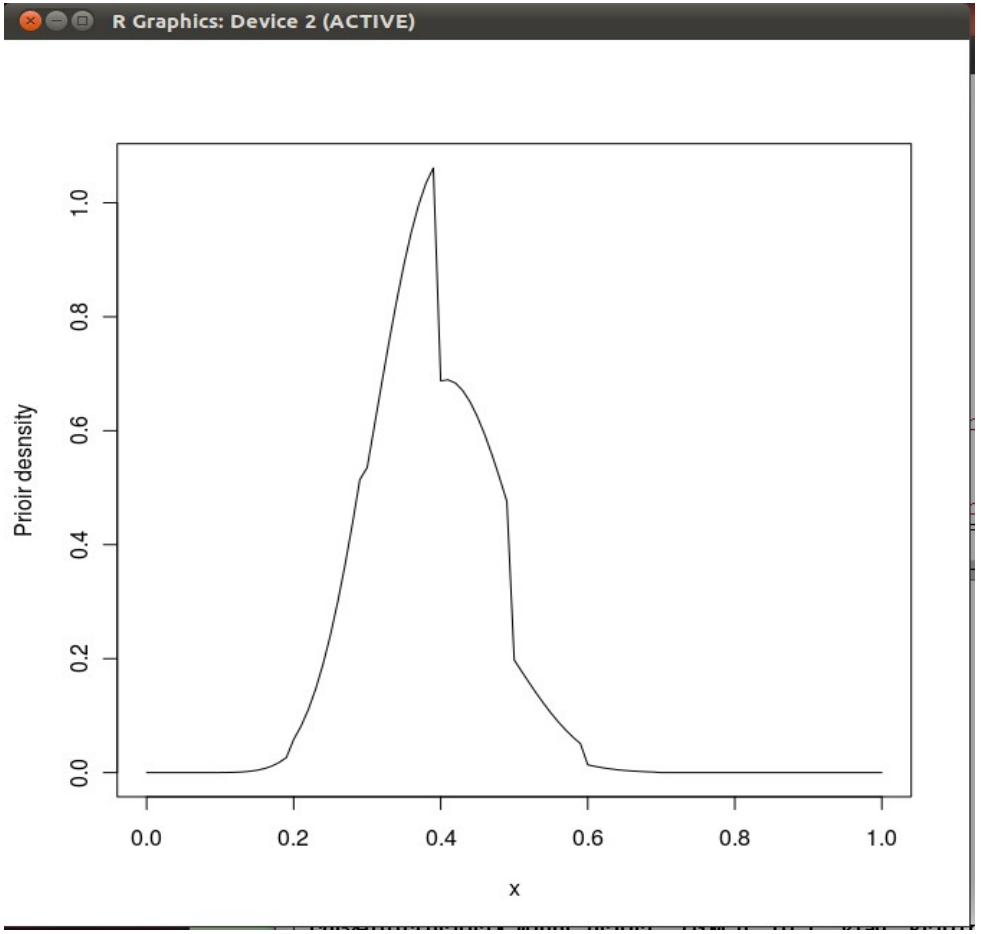
ヒストグラム事前分布(2)



事後分布のグラフ

> curve(histprior(x,midpt,prior) * dbeta(x, s+1, f+1), from=0, to=1, ylab="Prioir desnsity")

$$g(data|p) \propto p^s(1-p)^f = (B(s+1, f+1))'$$



予測

P の推定の後に、m = 20 人のうち、何人が
睡眠時間 8時間以上かを推定したい

y 人、y は確率変数、分布をもつ f(y)

$$f(y) = \int f(y|p)g(p)dp$$

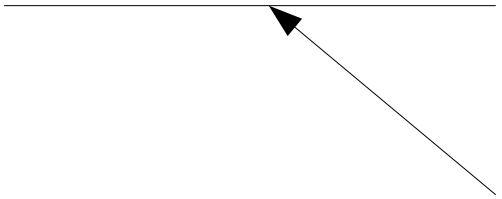
g(p) が事前密度なら f(y) は事前予測密度
g(p) が事後密度なら f(y) は事後予測密度

離散事前分布に対する f(y)（1）

$g(p_i)$　は以前与えたもの

$$f(y|p) =_m C_y p^y (1-p)^{m-y}$$

$$f(y) = \sum f(y|p_i) g(p_i)$$



この部分の計算 pdiscp()

離散事前分布に対する f(y)　(2)

```
> m <- 20; ys <- 0:20
> pred <- pdiscp(p,prior,m,ys)
> round(cbind(0:20, pred), 3)
          pred
 [1,]  0 0.020
 [2,]  1 0.044
 [3,]  2 0.069
 [4,]  3 0.092
 [5,]  4 0.106
 [6,]  5 0.112
 [7,]  6 0.110
 [8,]  7 0.102
 [9,]  8 0.089
[10,]  9 0.074
[11,] 10 0.059
[12,] 11 0.044
[13,] 12 0.031
[14,] 13 0.021
[15,] 14 0.013
[16,] 15 0.007
[17,] 16 0.004
[18,] 17 0.002
[19,] 18 0.001
[20,] 19 0.000
[21,] 20 0.000
```

4、5、6、7 が高い確率
4 ～ 7 人を予想

ベータ事前分布に対する f(y)

$$f(y) = \int f(y|p)g(p)dp =_m C_y \frac{B(a+y, b+m-y)}{B(a,b)}$$

この部分の計算 pbetap()

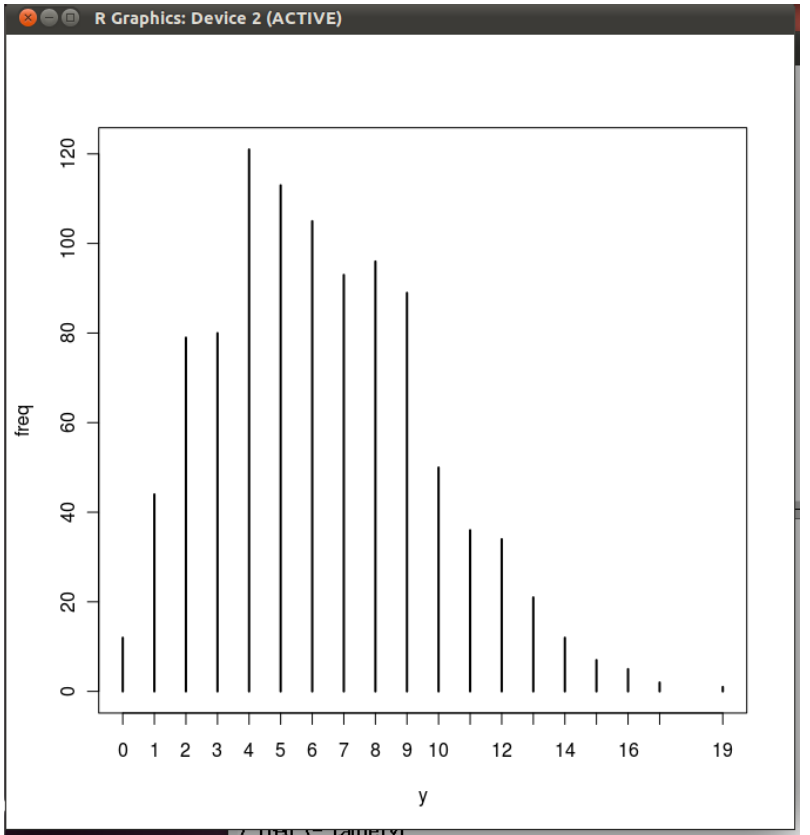
```
> pred <- pbetap(c(3.26, 7.19), m, ys)
> round(cbind(0:20, pred), 3)
          pred
 [1,]  0 0.018
 [2,]  1 0.045
 [3,]  2 0.072
 [4,]  3 0.095
 [5,]  4 0.108
 [6,]  5 0.114
 [7,]  6 0.111
 [8,]  7 0.102
 [9,]  8 0.088
....
[21,] 20 0.000
```

シミュレーションによる f(y)

(任意の)事前分布に従った p を生成
f(y|p) に従った y を生成
生成された y の分布から f(y) を推定

例) ベータ事前密度 B(3.26, 7.19) を利用

```
> p <- rbeta(1000, 3.26, 7.19)
> y <- rbinom(1000, 20, p)
> freq <- table(y)
> plot(freq)
```



f(y) から信頼区間の推定

(テキストの記述は冗長なので、概略だけ示す)

先のシミュレーションによる f(y) は頻度のグラフ、
全体の頻度で割って、確率が求まる、これが f(y)

確率の累積が 0.9 となるような区間を求めと
1 から 11 となる