Rで学ぶベイズ統計学入門

第4章 複数パラメータモデル

4．4 生物検定実験
4．5 二つの割合を比較する

茨城大学工学部情報工学科
國井慎也

生物検定実験

- 生物検定実験
  - 様々な濃度の化合物を動物群に投与し，その毒性を検定する実験

- 化合物の濃度，動物の数，集団それぞれに対する死亡数を記録したGelman[2003]のデータを使用する。

生物検定実験（2）

- 各集団の識別記号をiとすると

  x_i：投与した化合物の濃度

  n_i：動物の数

  y_i：動物の死亡数

  p_i：死亡率

- y_iは二項分布 binomial(n_i,p_i) に従うと仮定する

$$p(y_i) = binomial(n_i, p_i) = \binom{y_i}{n_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

- P_iはロジスティックモデルに従う

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 \qquad p_i = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

βの尤度関数

- 未知パラメータ β に関する尤度関数

$$L(\beta_0, \beta_1) \propto \prod_{i=1}^{4} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

ただし、$p_i = \dfrac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$

- まず，β の推定には最尤推定(非ベイズ)を用いる
  - ここでは，Rに用意されている一般化線形回帰モデルの最尤推定を行うglm関数を用いる

Rでの実行（最尤推定）

```
x <- c(-0.86,-0.3,-0.05,0.73)

n <- c(5,5,5,5)

y <- c(0,1,3,5)

data <- cbind(x,n,y)

response <- cbind(y,n-y)

results <-
glm(response~x,family=binomial)

summary(results)
```

```
Call:
glm(formula = response ~ x, family = binomial)

Deviance Residuals:
      1       2       3       4
-0.17236  0.08133 -0.05869  0.12237

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8466     1.0191   0.831    0.406
x             7.7488     4.8728   1.590    0.112

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15.791412  on 3  degrees of freedom
Residual deviance:  0.054742  on 2  degrees of freedom
AIC: 7.9648

Number of Fisher Scoring iterations: 7
```

.

$\beta_0 = 0.8466$ ， $\beta_1 = 7.7488$ と推定される

事前情報

- 次に、ベイズでの推定の準備として、ここでユーザには回帰のパラメータに事前の次の確信を持つものとする。

  投与量レベルx_L＝-0.7のときの、死亡確率p_Lの中央値と90%分位点は、それぞれ、0.2と0.5とする。

  投与量レベルx_H＝0.6のときの、死亡確率p_Hの中央値と90%分位点は、それぞれ、0.8と0.98とする。

- これらの情報に適合するベータ分布は以下で求められる。

  > beta.select(list(p=.5,x=.2),list(p=.9,x=.5))

  [1] 1.12 3.56

  > beta.select(list(p=.5,x=.8),list(p=.9,x=.98))

  [1] 2.10 0.74

- **結果、それぞれbeta(1.12,5.56)とbeta(2.10,0.74)に適合する。**

事前分布

- p_Lとp_Hが互いに独立であるとすると、結合事前分布は、

$$g(p_L, p_H) \propto p_L^{1.12-1}(1-p_L)^{3.56-1} p_H^{2.10-1}(1-p_H)^{0.74-1}$$

- この(p_L,p_H)から(β_0,β_1)に変換するための以下の式を使う。

$$p_L = \frac{\exp(\beta_0+\beta_1 x_L)}{1+\exp(\beta_0+\beta_1 x_L)} \qquad p_H = \frac{\exp(\beta_0+\beta_1 x_H)}{1+\exp(\beta_0+\beta_1 x_H)}$$

- すると、これより導かれる事前分布は次のようになる

$$g(\beta_0, \beta_1) \propto p_L^{1.12}(1-p_L)^{3.56} p_H^{2.10}(1-p_H)^{0.74}$$

事後分布

$$g(\beta_0, \beta_1) \propto p_L^{1.12}(1-p_L)^{3.56}\, p_H^{2.10}(1-p_H)^{0.74}$$

- この事前分布と尤度は同じ形であり、ベータ分布のパラメータは、事前実験での死亡数と生存数とみなすことができる。

- この事前分布と尤度から次の事後分布が得られる。

$$L(\beta_0, \beta_1 | y) \propto \prod_{i=1}^{6} p_i^{y_i}(1-p_i)^{n_i - y_i}$$

事後分布を用いた分析

次に，事後分布を用いた分析を行う

LearnBayesで用意されている関数を用いる

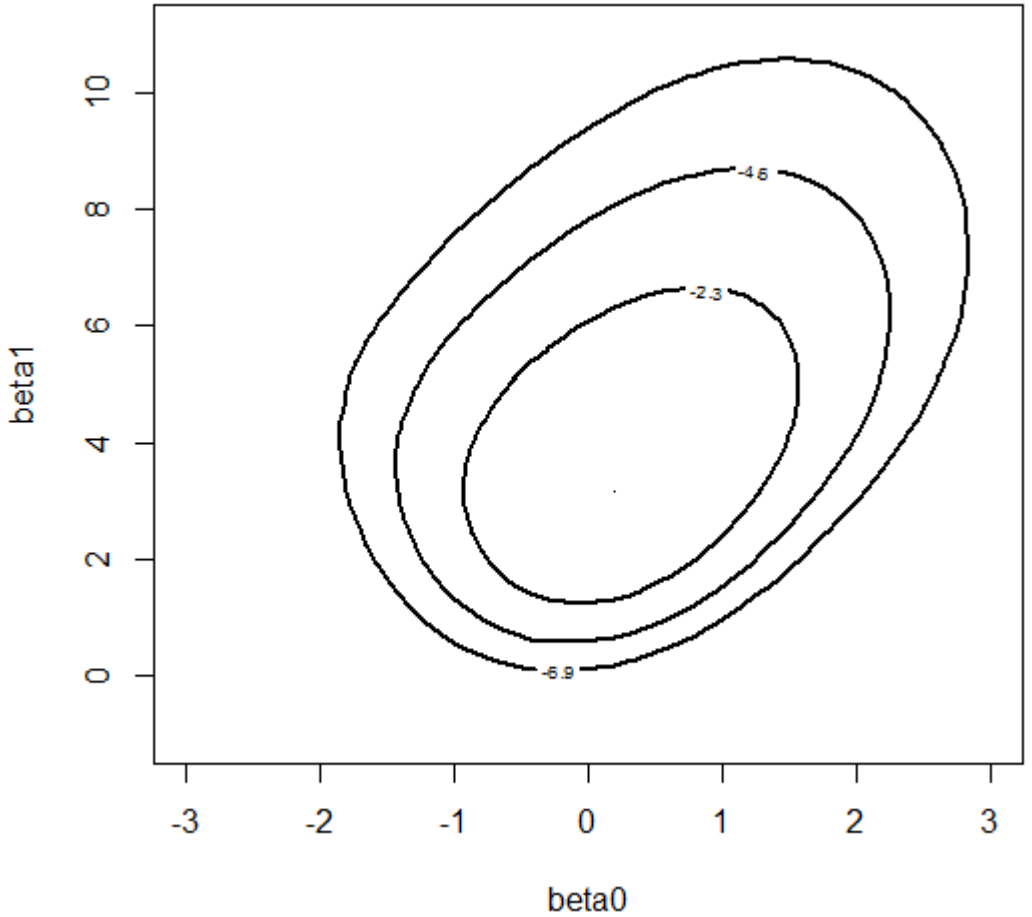
logisticpost: ロジスティックモデルの事後分布を計算する関数

mycontour: 与えられた2パラメータ密度の等高線を描く関数

simcontour: 与えられた2パラメータについてランダムな標本をシミュレーションする関数

(β_0,β_1)の事後分布の等高線

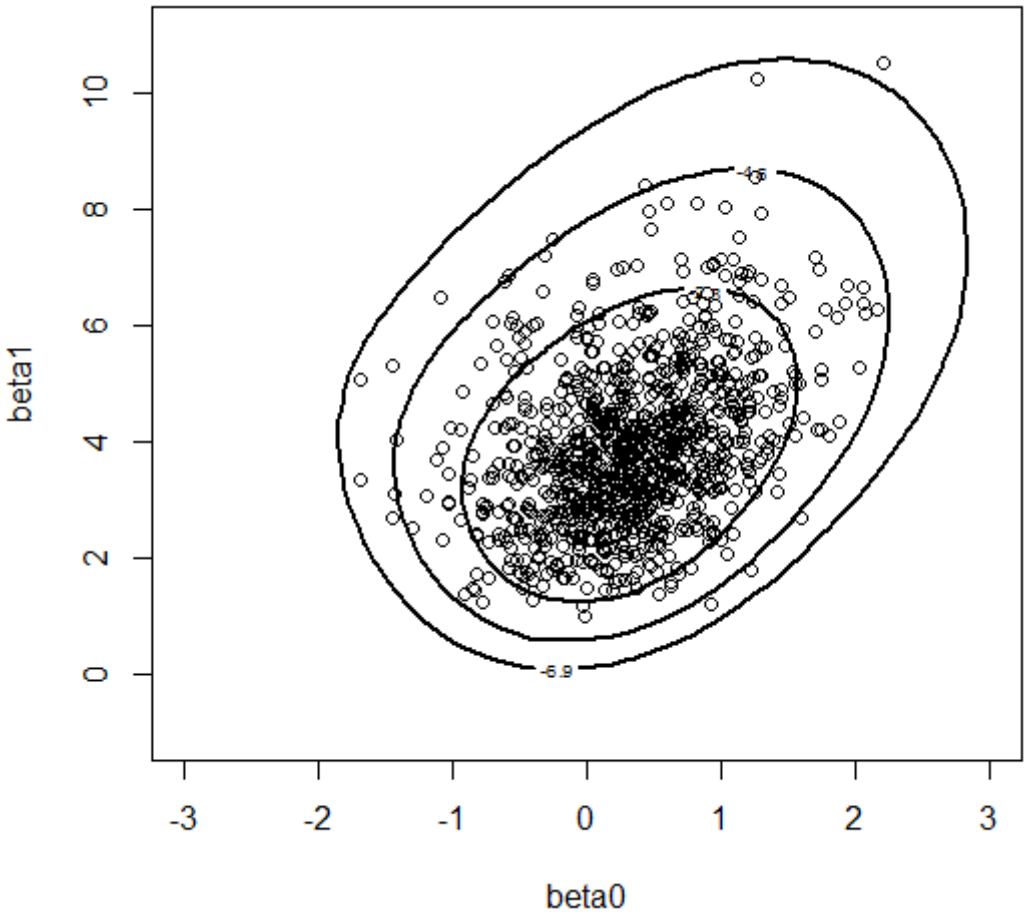
```
> data <- cbind(x,n,y)
> prior <- rbind(c(-0.7,4.68,1.12),c(0.6,2.84,2.10))
> data.new <-rbind(data,prior)
> mycontour(logisticpost,c(-3,3,-1,11),data.new,
+ xlab="beta0", ylab="beta1")
```



- 生物検定事例における(β_0,β_1)の事後分布の等高線.等高線は、モードの高さの10%,1%,0.1%で描かれている.

事後密度から(β_0,β_1)の値をシミュレーションする

```
> s <- simcontour(logisticpost,c(-3,3,-1,11),data.new,1000)
> points(s)
```

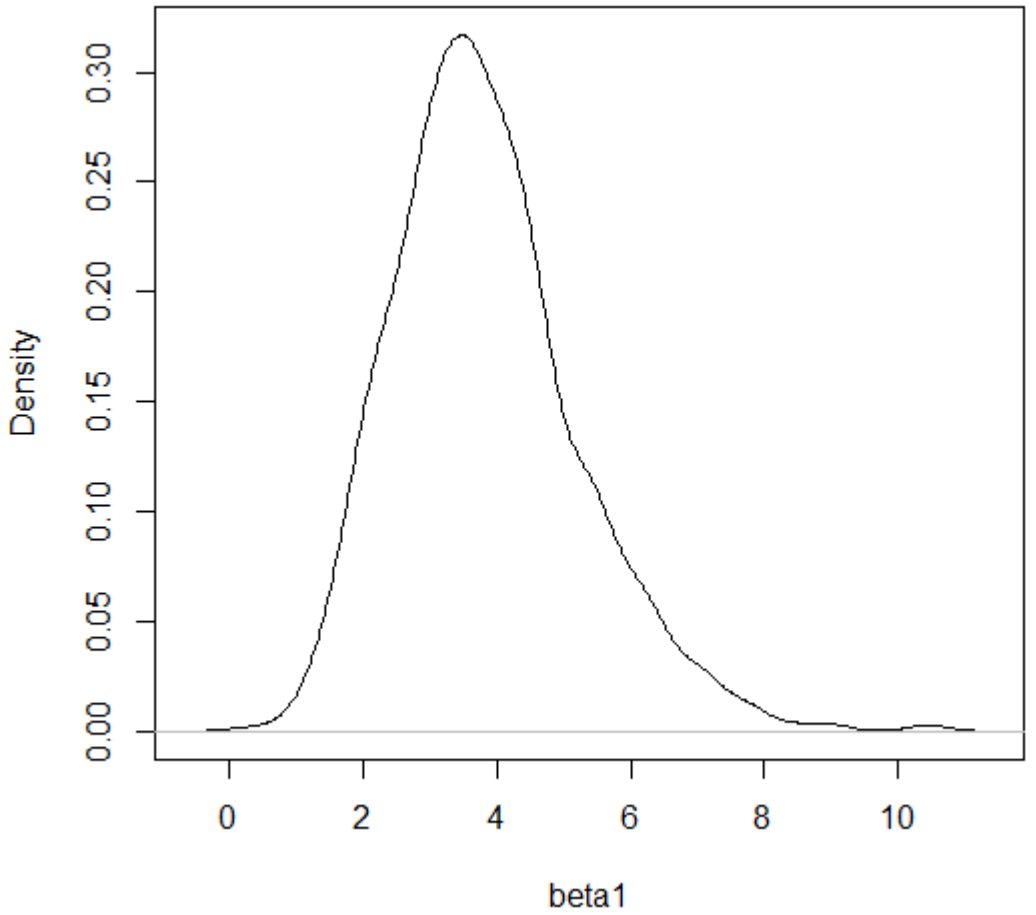


- 生物検定事例における(β_0,β_1)の事後分布の等高線図。この事後分布からシミュレーションしたランダムな標本を等高線図に重ねている。

β_1の密度質量

```
> plot(density(s$y),xlab="beta1")
```

- β_1の密度質量の範囲は正の位置にあり、投薬量のレベルが上がると死亡確率も上がるという証拠が得られる。



LD-50

- LD-50

  死亡確率が50%になる投薬量xの値。

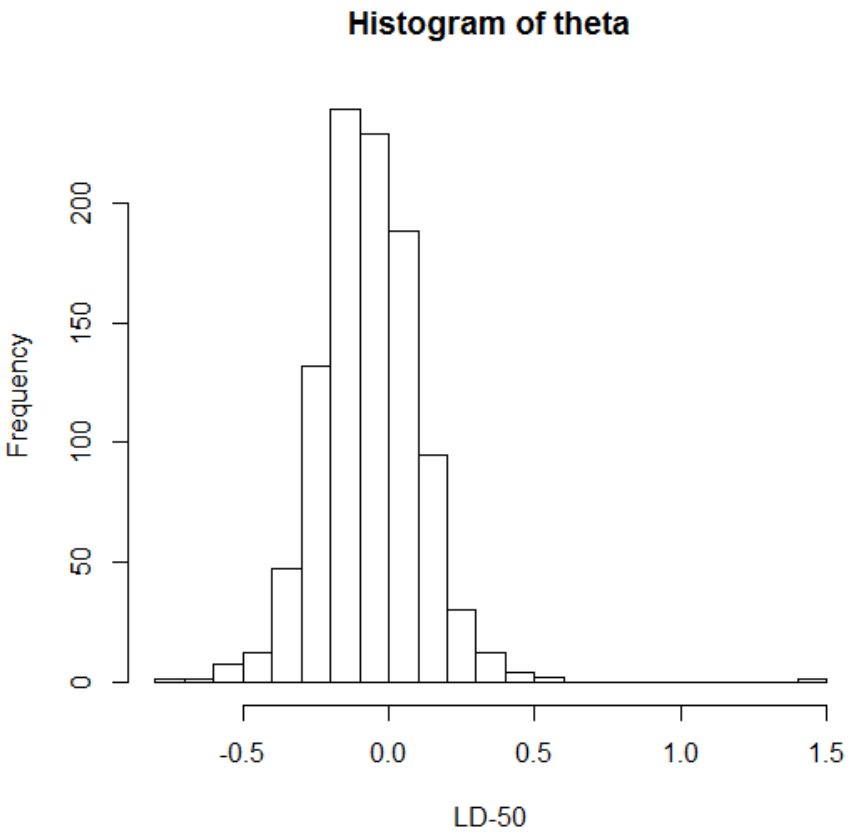
  LD-50の値は、$\theta=-\beta_0/\beta_1$で与えられる

- LD-50の値を求めるためには、シミュレーションした(β_0,β_1)からθを計算すればよい

LD-50の推定

```
> theta <- -s$x/s$y
> hist(theta,xlab="LD-50",breaks = 20)
```

- グラフを見ると、β_1のヒストグラムLD-50の推定は難しく、この事後パラメータの幅は広くなる。



LD-50の95%の信頼区間

- 95%の信頼区間は、求めると次のようになる。

```
> quantile(theta,c(0.025,.975))
      2.5%      97.5%
-0.3542899  0.5061084
```

- θが区間(-0.354,0.506)に含まれる確率は95%である。

二つの割合を比較する

- ２つの独立した二項分布に関する比率を比較したい

  y1：二項分布（n1,p1）

  y2：二項分布（n2,p2）

  仮説1：p1＞p2、仮説２：p1<p2
- p1が0.8に近いという証拠を得た場合,p2もおそらく0.8に近いのではないかと確信を与える.

  よって、これらの事前分布に独立な分布を用いるより、依存関係を持つ分布を用いる方が適切。

- Hawardによる従属事前分布

従属事前分布

- 割合は、実数値のロジットパラメータに変換される。

$$\theta_1 = \log \frac{p_1}{1-p_1} \qquad \theta_2 = \log \frac{p_2}{1-p_2}$$

θ１が与えられると、θ2は平均θ１で標準偏差σの正規分布に従うと仮定する。

- 従属事前分布は

$$g(p_1, p_2) \propto e^{-(\frac{1}{2})u^2} p_1^{\alpha-1}(1-p_1)^{\beta-1} p_2^{\gamma-1}(1-p_2)^{\delta-1}$$

ここで、

$$u = \frac{1}{\sigma}(\theta_1 - \theta_2)$$

- この従属事前分布は、パラメータ($\alpha, \beta, \gamma, \delta, \sigma$)で表せる。
- 最初のパラメータは、p1とp2の位置の確信を表し、σは依存関係の強さを表している。

従属事前分布

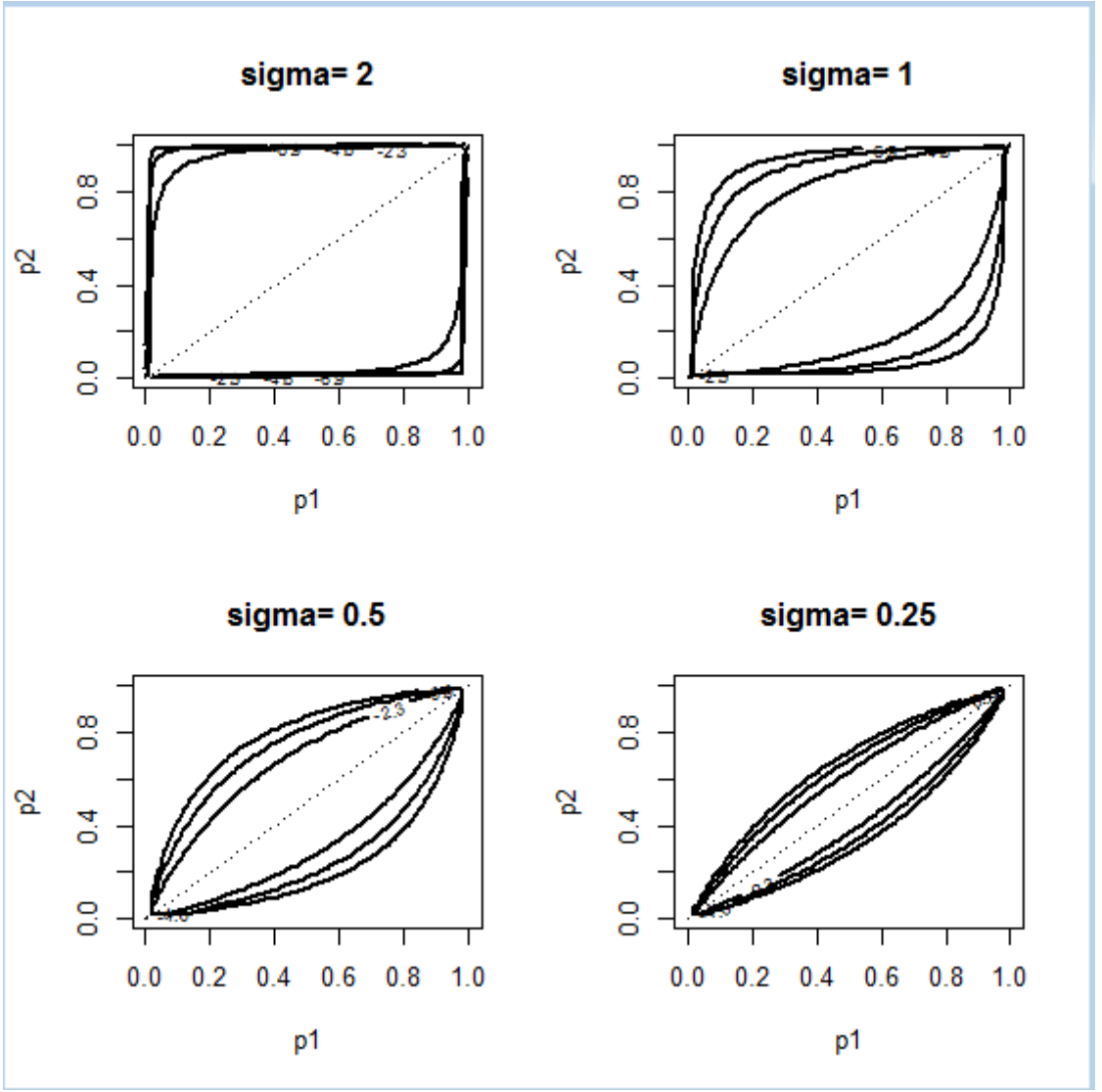
- 従属事前分布の対数は、howardprior()関数で求められる。

  α=β=γ=δ=1とする。

```
 sigma <- c(2,1,.5,.25)
> plo <- .0001; phi<-.9999
> par(mfrow=c(2,2))
>  for(i in 1:4){
+  mycontour(howardprior,c(plo,phi,plo,phi),
+ c(1,1,1,1,sigma[i]),
+ main =   paste("sigma=",as.character(sigma[i])),
+ xlab = "p1",ylab = "p2")
+ }
```

従属事前分布の等高線

- σの値が0に近づくにつれて、事前分布の質量は、二つの割合が等しいときの直線に近づく。



パラメータの更新

- この二つの二項分布から、それぞれy_1, y_2の観測数を得るときの、尤度は以下のようになる。

$$L(p_1, p_2) = p_1^{y_1}(1-p_1)^{n_1-y_1} p_2^{y_2}(1-p_2)^{n_2-y_2}$$

- 尤度と事前分布を結合すると、パラメータが次のように更新される。

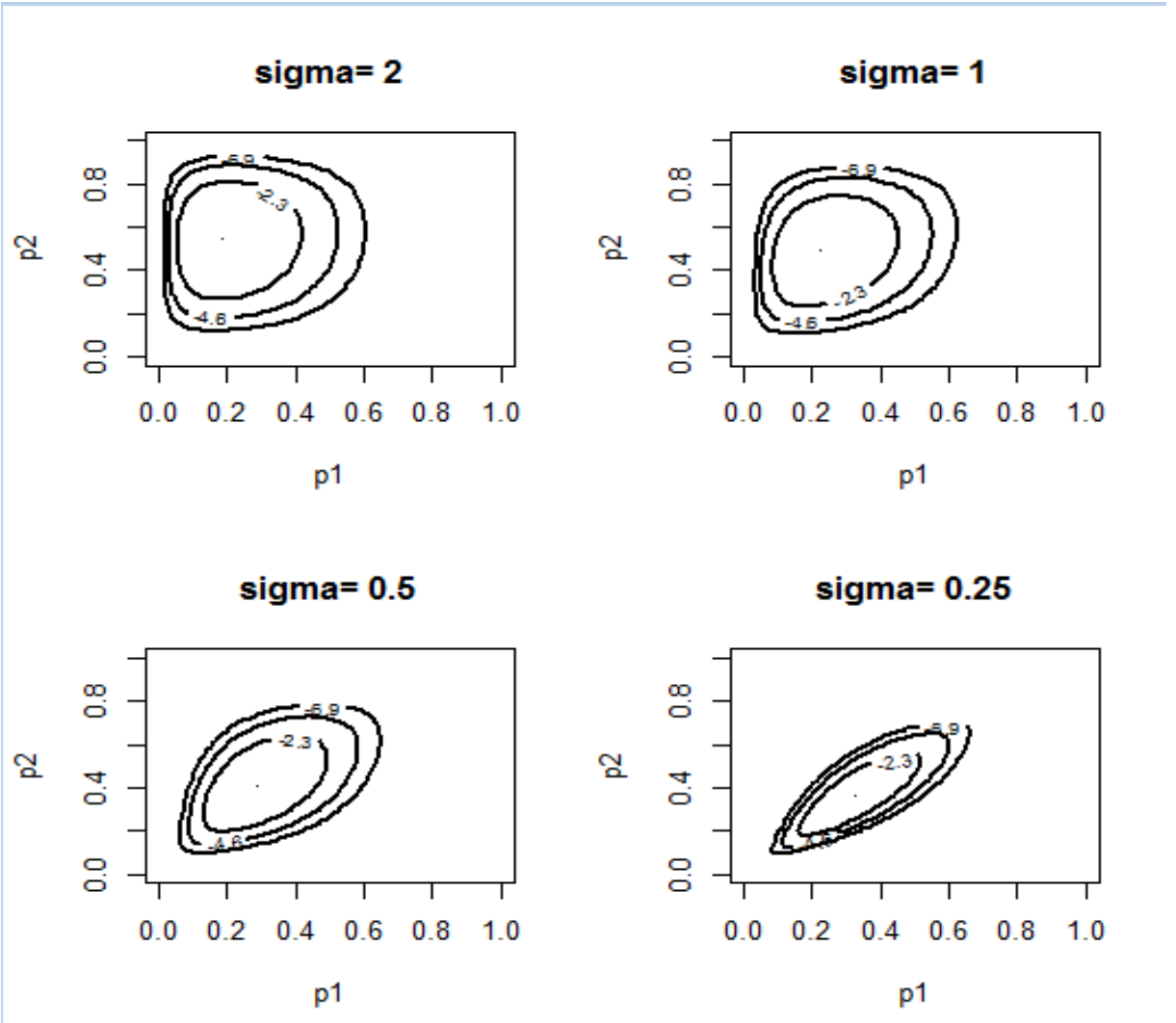
$$(\alpha+y_1, \beta+n_1-y_1, \gamma+y_2, \delta+n_2-y_2, \sigma)$$

事後分布の等高線図

- 次のデータセットを使う

	成功	失敗	合計
標本1	3	15	18
標本2	7	5	12
合計	10	20	30

```
 sigma <- c(2,1,.5,.25)
> par(mfrow=c(2,2))
>  for(i in 1:4){
+  mycontour(howardprior,c(plo,phi,plo,phi),
+ c(1+3,1+15,1+7,1+5,sigma[i]),
+  main =
paste("sigma=",as.character(sigma[i])),
+ xlab = "p1",ylab = "p2")
+ }
```



仮説の検定(1)

- 仮説H1の検定

  パラメータ空間の事後確率を計算する。

  最初に、simcontour()を使い、(p_1,p_2)の事後分布から標本を求める。次に、p_1>p_2である割合を求める

- 以下にσ＝２のコードを示す

```
s<-simcontour(howardprior,c(plo,phi,plo,phi),
+ c(1+3,1+15,1+7,1+5,2),1000)
> sum(s$x>s$y)/1000
```

仮説の検定(2)

- 求めた事後分布の結果を以下に示す。

従属度パラメータ	P(p1>p2)
2	0.012
1	0.035
0.5	0.102
0.25	0.201

- 求めた事後確率は、二つの割合間の依存関係についての事前の確信に対して敏感である。