Rで学ぶベイズ統計学入門
第３章　１パラメーターモデル

茨城大学工学部
佐々木研究室
高木真

平均が既知で分散が未知の正規分布

- Gelman et al.(2003)が用いたアメリカンフットボールのスコアを例に道の分散を推定する。
- 試合の結果(得点と失点の差)とポイントスプレッドの差dを考える。

モデル分布

- n回の試合についてのポイントスプレッドの観測値の差が$d_1, \cdots, d_n$であるとする。
- これらの差が、平均0で未知分散σ^2の正規分布に従うとすると、このとき尤度関数は

$$L(\sigma^2) \propto (\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^{n} d_i^{\,2} /(2\sigma^2)\}(\sigma^2 > 0)$$

となる。

事後密度

- 分散に無情報事前密度 $P(\sigma^2) \propto 1/\sigma^2$ を割り当てると、σ^2 の事後密度は

$$g(\sigma^2 \mid data) \propto (\sigma^2) - n/2 - 1 \exp\{-v/(2\sigma^2)\}$$

となる。

ここで、$v = \sum_{i=1}^{n} d_i^{\ 2}$ である。

- 精度パラメータをP=1/σ²と仮定すると、PはU/vとして分布することを示す。

ここでUは自由度nのカイ二乗分布にしたがう。

Rによる実行(1)

- Rによって事後分布を算出する。

- まず、LearnBayesパッケージのfootballscoresデータファイルを読み込む。
  - favorite 強いとみられているチームのスコア
  - underdog 弱いとみられているチームのスコア
  - spread 公表されたポイント・スプレッド

Rによる実行(2)

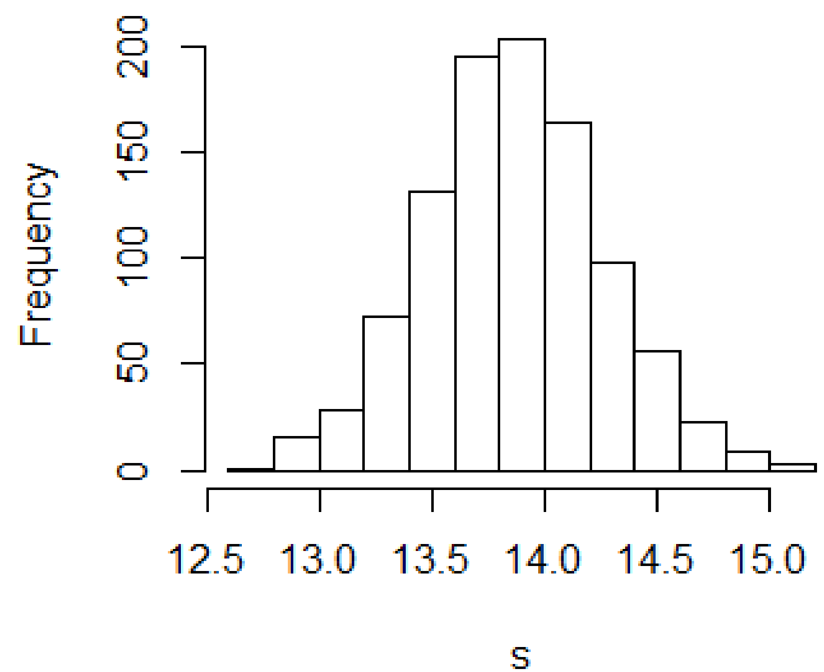
- 差の変数dの計算
  - n 標本サイズ
  - v 差の平方和

```
> data(footballscores)
> attach(footballscores)
> d <- favorite – underdog – spread
> v <- sum(d^2)
```

Rによる実行(3)

- 標準偏差σの事後分布から1000個の値をシミュレーションする。

```
>p <- rchisq(1000,n)/v
>s <- sqrt(1/p)
>hist(s, main = “”)
```



Rによる実行(4)

- quantili()関数を使い、2.5%、50%、97.5%分位点を求める。

```
> quantile(s, probs = c(0.025, 0.5, 0.975))
    2.5%        50%       97.5%
13.17012   13.85135   14.56599
```

以上より、事後分布の中央値が13.85がσの点推定値となり、両端の分位点(13.2,14.6)がσの95%確率区間となる。

心臓移植手術の死亡率の推定(1)

- ある病院での心臓移植手術の成功率を調べるという問題を考える。
- わかっているデータ
  - e 移植手術数
  - y 執刀後30日以内の死亡数
  - $e\lambda$ 死亡数の期待値

ここで死亡数yは平均$e\lambda$のポアソン分布に従うと想定される

心臓移植手術の死亡率の推定(2)

- 目標

単位曝露あたりの死亡率の推定

- λの通常の推定値は最尤推定値の$\hat{\lambda}$ =y/eであるが、死亡数yが0に近い場合、誤差が大きくなることがある。

死亡率の高さについての事前知識を取り入れたベイズ推定値を利用することが望ましい。

事前分布

- 事前分布として選択して便利なのが、次の式で表現されるガンマ密度gamma(α,β)の一つである。

$$p(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda) \quad (\lambda>0)$$

- 事前情報として調査対象と死亡率が同じであろう病院でのデータを用いる。
- z_j 10の病院での死亡数(j=1,･･･,10)
- o_j 曝露数

z_jは平均$o_j\lambda$のポアソン分布に従うものとする。

事前分布(2)

- 仮にλに標準的な無情報事前分布 $p(\lambda) \propto \lambda^{-1}$ を割り当てるとλの分布は次のように更新される。

$$p(\lambda) \propto \lambda \sum_{j=1}^{10} z_j - 1 \exp(-(\sum_{j=1}^{10} o_j)\lambda)$$

この情報を使うとλの事前分布はガンマ分布 gamma(α,β)である。 (α = $\sum_{j=1}^{10} z_j$,β = $\sum_{j=1}^{10} o_j$)

ここでの事例に当てはめると以下のようになる。

$\sum_{j=1}^{10} z_j$ =16, $\sum_{j=1}^{10} o_j$ = 15174

Rによる実行(1)

- ベイズ法による計算
  - alpha,beta ガンマ事前分布のパラメータ
  - ex 曝露数

```
> alpha <- 16; beta <- 15174
>yobs <- 1; ex <-66
> y <- 0:10
> lam <- alpha/beta
>py <- dpois(y,lam * ex) * dgamma(lam, shape =
+ alpha,rate = beta)/dgamma(lam, shape = alpha +y,
+ rate = beta + ex)
```

Rによる実行(2)

```
>cbind(y, round(py,3))
 y
 [1,]  0 0.933
 [2,]  1 0.065
 [3,]  2 0.002
 [4,]  3 0.000
 [5,]  4 0.000
 [6,]  5 0.000
 [7,]  6 0.000
 [8,]  7 0.000
 [9,]  8 0.000
[10,]  9 0.000
[11,] 10 0.000
```

Rによる実行(3)

- λの事後分布はガンマ密度から1000個の値をシミュレーションすることで要約される。

```
>lambdaA <- rgamma(1000,shape = alpha + yobs,
  rate = beta +ex)
```

```
> curve(dgamma(x, shape = alpha, rate = beta),
  add = TRUE)
```

Rによる実行(4)

- e=1767,y=4の場合

```
> ex2 <-1767; yobs2 <-4
> py <- dpois(y,lam *ex2)*dgamma(lam, shape =
  alpha,
+ rate = beta)/dgamma(lam, shape = alpha + y,
+ rate = beta + ex2)
```

Rによる実行(5)

```
> cbind(y, round(py,3))
      y
 [1,]  0 0.172
 [2,]  1 0.286
 [3,]  2 0.254
 [4,]  3 0.159
 [5,]  4 0.079
 [6,]  5 0.033
 [7,]  6 0.012
 [8,]  7 0.004
 [9,]  8 0.001
[10,]  9 0.000
[11,] 10 0.000
```

Rによる実行(6)

- 同様にガンマ密度から1000個の値をシミュレーションする

```
> lambdaB <- rgamma(1000, shape = alpha + yobs, rate = beta + ex)
> plot(density(lambdaB),main = "HOSPITAL B",
+ xlab = "lambdaB", lwd = 3)
> curve(dgamma(x, shape = alpha, rate = beta),add = TRUE)
```

事前密度と事後密度

- 二つの病院における心臓移植手術の死亡率の事前密度と事後密度