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

4.1 はじめに

4.2 パラメータが二つとも未知の正規データ

4.3 多項モデル

茨城大学工学部

佐々木稔

はじめに

- 未知パラメータが複数ある場合のベイズモデル
  - 正規母集団のパラメータ（4.2節）
  - 多項分布モデルのパラメータ（4.3節）
  - ロジスティック単回帰モデルのパラメータ（4.4節）
  - 二項分布に従うふたつの割合を比較（4.5節）

パラメータが二つとも未知の正規データ

- 平均と分散が未知の正規母集団
  - 事後分布をベイズ推定により計算
- 20代男性ランナー20人の完走時間の分布
  - N(μ, σ) に従うランダム標本と仮定
  - さらに、標準的な無情報事前分布を仮定
  - 平均と分散の事後分布
    - n は標本サイズ、$\bar{y}$ は標本平均、$S = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$g(\mu, \sigma^2 \mid y) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2} \left(S + n(\mu - \bar{y})^2 \right) \right)$$

正規データの事後確率密度分布

- 平均、分散の事後分布
  - 分散が分かっている条件での平均の事後分布
    - 正規分布 $N\left(\bar{y}, \sigma/\sqrt{n}\right)$ に従う
  - 分散の周辺事後分布は $S\chi_{n-1}^{-2}$ に従う
    - 自由度 n-1 の逆カイ二乗分布

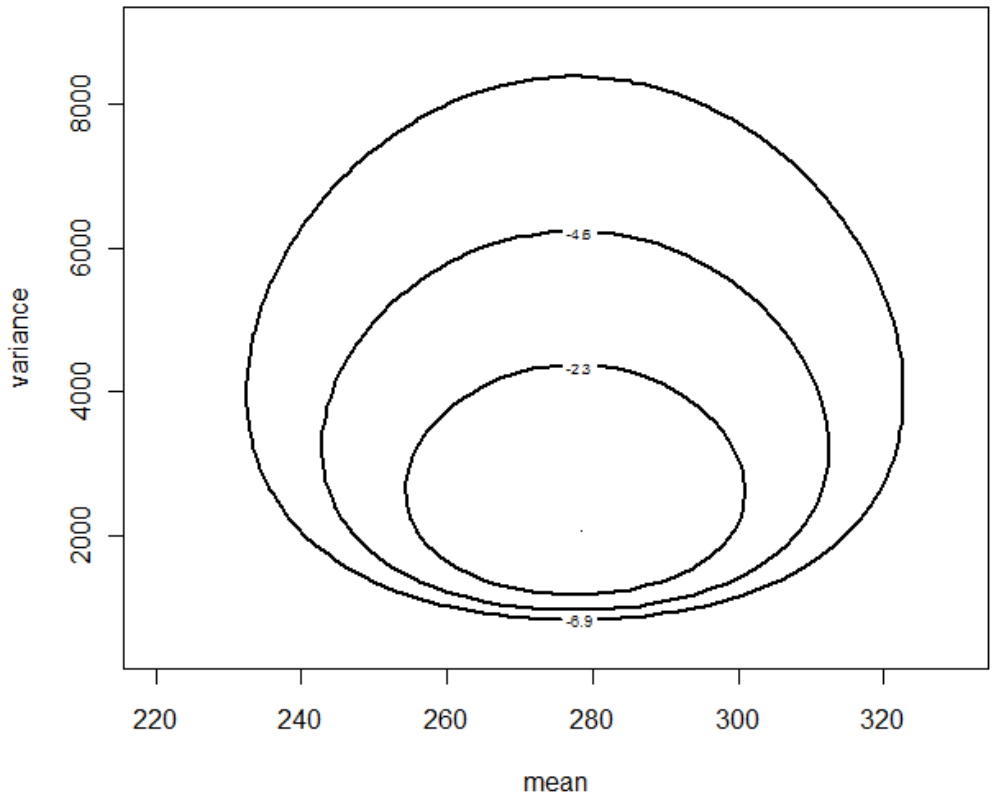
Rによる事後確率密度分布の計算

- 事後確率密度分布の等高線図を描く
  - データ marathontimes の読み込み
  - normchi2post()関数
    - 平均と分散の事後密度の対数を計算
  - mycontour()関数
    - 等高線図を描くための関数

```
> data(marathontimes)
> attach(marathontimes)
> d <- mycontour(normchi2post, c(220, 330, 500, 9000), time, xlab = "mean", ylab = "variance")
```

平均と分散の事後確率密度分布

- 対数密度関数を等高線図としてプロット
  - 事後確率密度の最大値の10%、1%、0.1%



シミュレーションによる推定

- 平均、分散をシミュレーションにより求める
  - $S\chi_{n-1}^{-2}$ の分布から分散を計算する
  - 正規分布 $N\left(\bar{y}, \sigma/\sqrt{n}\right)$ から平均を計算する
- 方法
  - rchisq()関数でカイ二乗分布から1000個抽出
  - これらの値に対して、分散を求める
  - さらに各値に対し、rnorm()関数で平均値を計算
  - 等高線図に点を上書き

シミュレーションのRコマンド

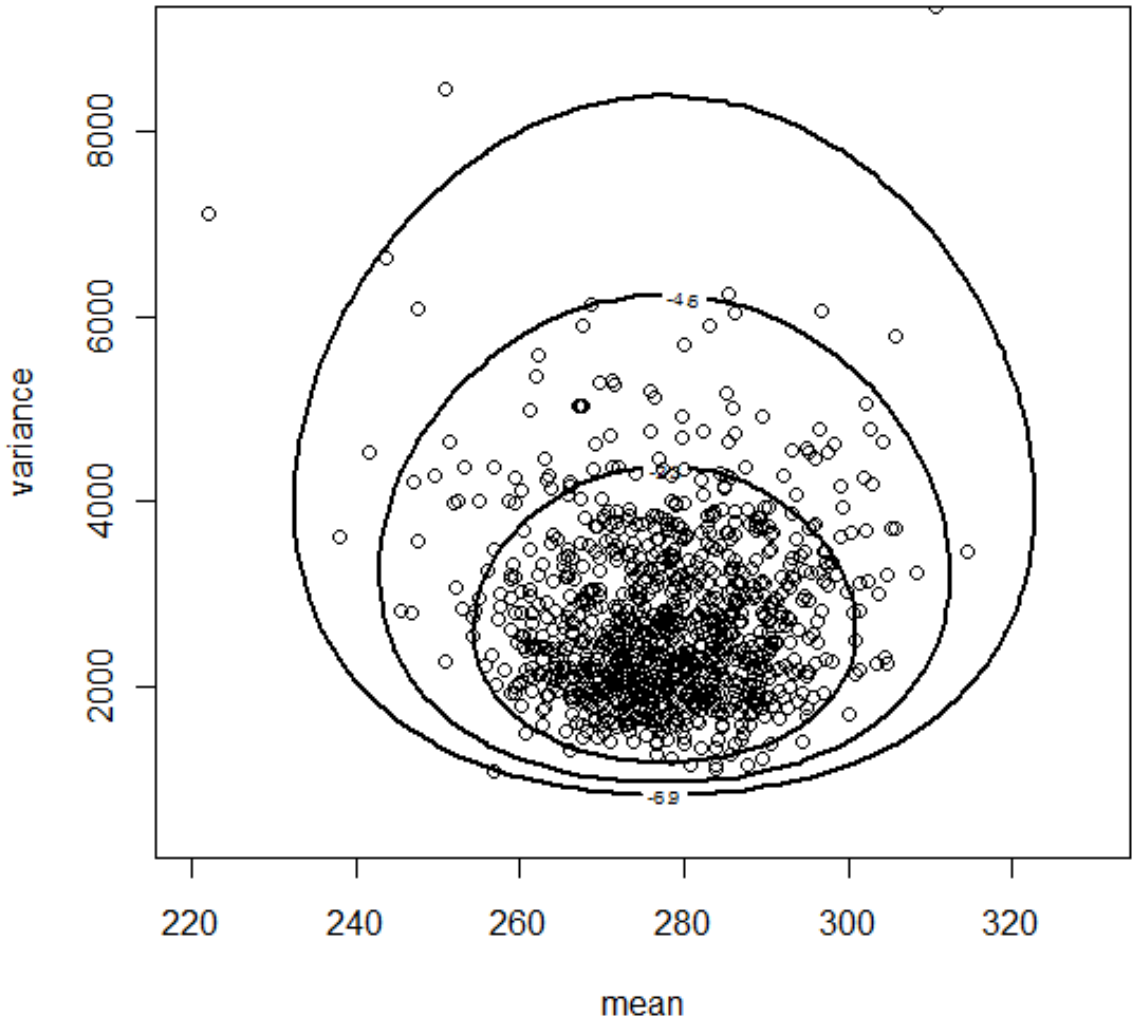
- コマンド

```
> S <- sum((time - mean(time))^2)
> n <- length(time)
> sigma2 <- S/rchisq(1000, n - 1)
> mu <- rnorm(1000, mean = mean(time), sd = sqrt(sigma2)/sqrt(n))
> par(new=T)
> points(mu, sigma2)
```

- normpostsim()関数でもシミュレーション可能

```
> param <- normpostsim(time, m=1000)
> points(param$mu, param$sigma2)
```

サンプリングによる平均と分散の表示



推定結果の信頼区間

- 平均 μ と標準偏差 σ の95%区間推定
  - quantile()関数を使う

```
> quantile(mu, c(0.025, 0.975))
   2.5%    97.5%
256.0119 299.9970
> quantile(sqrt(sigma2), c(0.025, 0.975))
   2.5%    97.5%
38.14511 70.96598
```

  - 平均完走時間の95%信頼区間は (256.0, 300.0)
  - 標準偏差の95%信頼区間は (38.1, 71.0)

多項モデル

- 1998年大統領選挙前の世論調査
  - ジョージ･ブッシュ支持　$y_1 = 727$（人）
  - マイケル･デュカキス支持　$y_2 = 583$（人）
  - その他の候補を支持、無回答　$y_3 = 137$（人）
- 標本サイズ n, 各確率 $\theta_1, \theta_2, \theta_3$ の多項分布

$$f(y_1, y_2, y_3; n; \theta_1, \theta_2, \theta_3) = \frac{n!}{y_1! y_2! y_3!} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3}$$

- 多項分布の共役事前分布
  - パラメータ (y_1+1, y_2+1, y_3+1) のディリクレ分布

事後分布のシミュレーション

- 事後分布はディリクレ分布

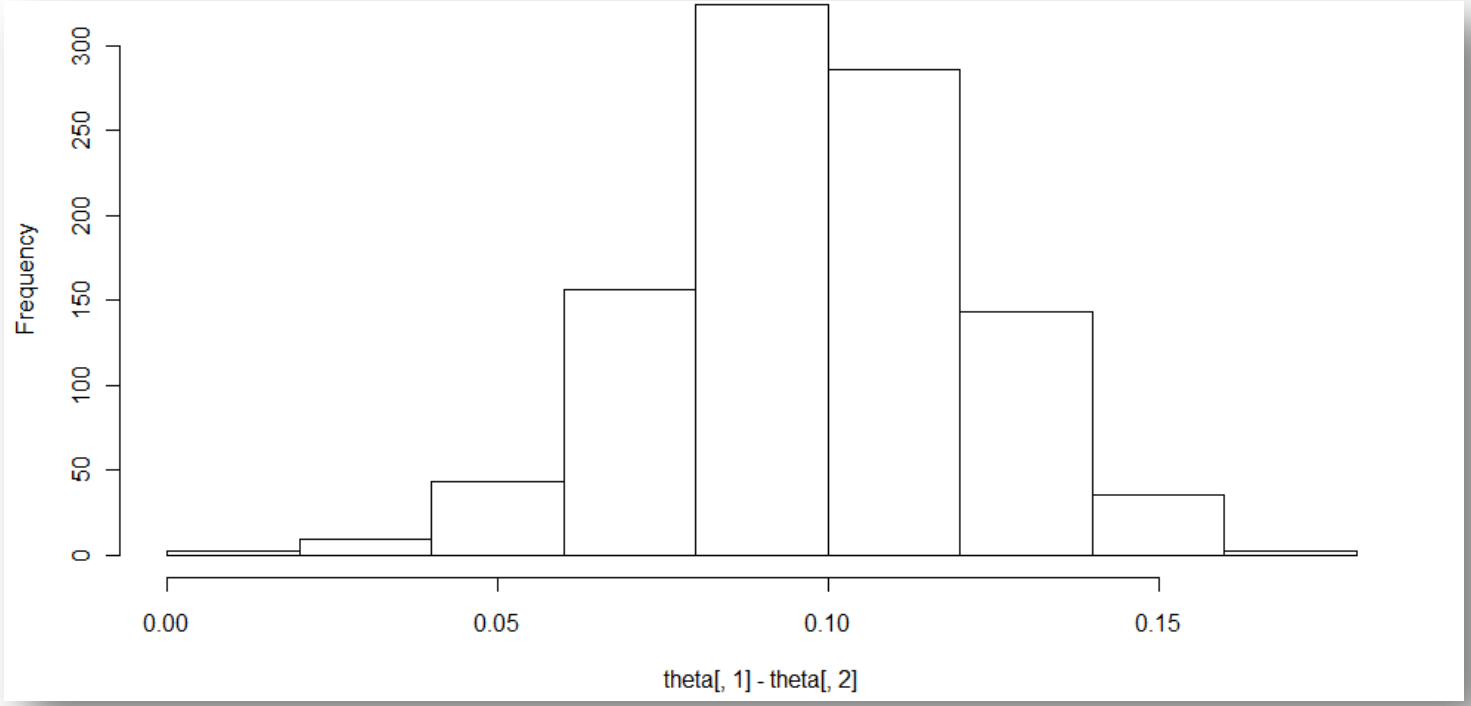
$$Dir(\theta_1, \theta_2, \theta_3 \mid data) = f(y_1, y_2, y_3; n; \theta_1, \theta_2, \theta_3) diri(\theta_1, \theta_2, \theta_3 \mid y_1 + 1, y_2 + 1, y_3 + 1)$$
$$\propto \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3}$$

  - 事後分布も (y_1+1, y_2+1, y_3+1) のディリクレ分布
- 事後分布のシミュレーション
  - rdirichlet()関数でディリクレ分布に従う乱数取得

```
> alpha <- c(728, 584, 138)
> theta <- rdirichlet(1000, alpha)
> hist(theta[,1] - theta[,2], main= "")
```

事後分布のシミュレーション結果

- θ1-θ2 のヒストグラム
  - 分布が正の値に限定
  - ブッシュがデュカキスを上回ることが分かる



選挙人投票数の獲得予測

- 2008年の大統領選挙（選挙人総計538人）
  - バラク・オバマ、ジョン・マケイン
  - 州に割り当てられた選挙人投票数を獲得
- オバマが獲得する選挙人投票数合計 EV_O を予測
  - j 番目の州で獲得する選挙人割合を θ_{Oj}, θ_{Mj}
  - j 番目の州の選挙人投票数 EV_j
    - $I()$関数は引数が真ならば 1、偽ならば 0 を返す

$$EV_o = \sum_{j=1}^{51} EV_j \times I\left(\theta_{O_j} > \theta_{M_j}\right)$$

獲得票の標本割合

- 各州での世論調査の標本数 500
  - オバマの獲得票の標本割合 q_{Oj}
  - マケインの獲得票の標本割合 q_{Mj}
- 事後分布
  - 一様事前分布を仮定
  - 選挙人割合と得票率は独立した事後分布
  - j 番目の州で候補者への投票割合
    - パラメータ $(500q_{Oj} + 1, 500q_{Mj} + 1, 500(1-q_{Oj}-q_{Mj}) + 1)$ のディリクレ分布

選挙人投票数の事後分布

- オバマの得票数の事後分布
  - 州ごとの割合を用いてシミュレーション
- データ
  - election.2008 データセット
    - 州毎のオバマへの投票割合 O.pct
    - 州毎のマケインへの投票割合 M.pct
    - 選挙人投票数 EV

```
> library(LearnBayes)
> data(election.2008)
> attach(election.2008)
```

事後確率の計算

- ディリクレ分布で標本抽出
  - *j* 番目の州で θ_{Oj} が θ_{Mj} を超える事後確率を計算
  - すべての州について、オバマが勝つ確率を計算
    - sapply()関数を使う

```
> prob.Obama <- function(j)
+ {
+     p <- rdirichlet(5000, 500 * c(M.pct[j], O.pct[j], 100 - M.pct[j] - O.pct[j])/100 + 1)
+     mean(p[, 2] > p[, 1])
+ }
> Obama.win.probs <- sapply(1:51, prob.Obama)
```

事後分布からシミュレーション標本の抽出

- オバマの選挙人獲得数を推定
  - 投票割合により各州の勝者を決める
  - 全体の獲得選挙人数を計算
- 1000回のシミュレーションを行う

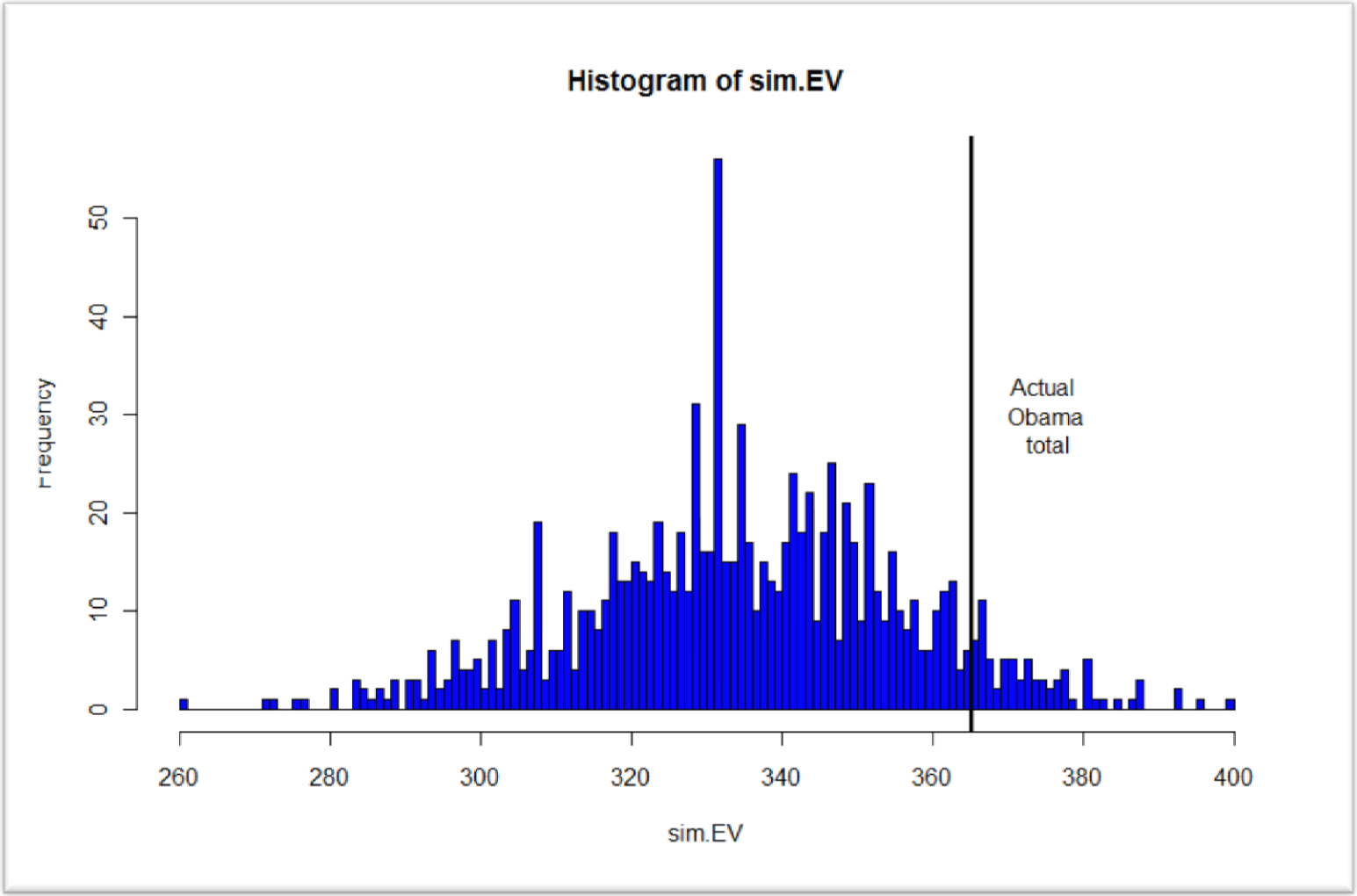
```
> sim.election <- function()
+ {
+     winner <- rbinom(51, 1, Obama.win.probs)
+     sum(EV * winner)
+ }
+ sim.EV <- replicate(1000, sim.election())
```

シミュレーション結果の可視化

- 推定した獲得数のヒストグラム
  - 実際に獲得した選挙人365も図に示す

```
> hist(sim.Ev, min(sim.EV): max(sim.EV), col="blue")
> abline(v=365,lwd=3)
> text(375,30,"Actual ¥n Obama ¥n total")
```

シミュレーション結果


Histogram of sim.EV
Frequency
0
10
20
30
40
50
260
280
300
320
340
360
380
400
sim.EV
Actual
Obama
total

結果の分析

- 実際の獲得数よりも少ない
  - 過半数（269人）は大幅に超えている
    - 勝つかどうかの推定はこのくらいの分析でも可能
  - データが少なすぎる
    - もう少し増やすと予測結果の改善が期待できる
  - 実際の選挙人獲得数は90%信頼区間の範囲内