2015年10月2日更新

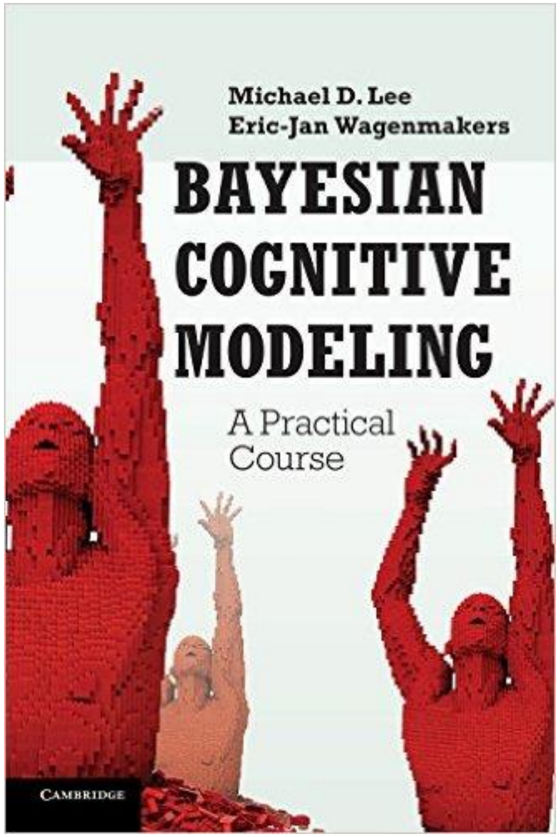
ベイズ推定の初歩
―二項分布を例に―

大阪大学大学院人間科学研究科
武藤拓之 (Hiroyuki MUTO)
http://kiso.hus.osaka-u.ac.jp/muto/

この資料はLee & Wagenmakers (2014) のChapter 3 (pp. 37-53), “Inferences with binomials” を武藤が解説したものです。

この資料について (1)

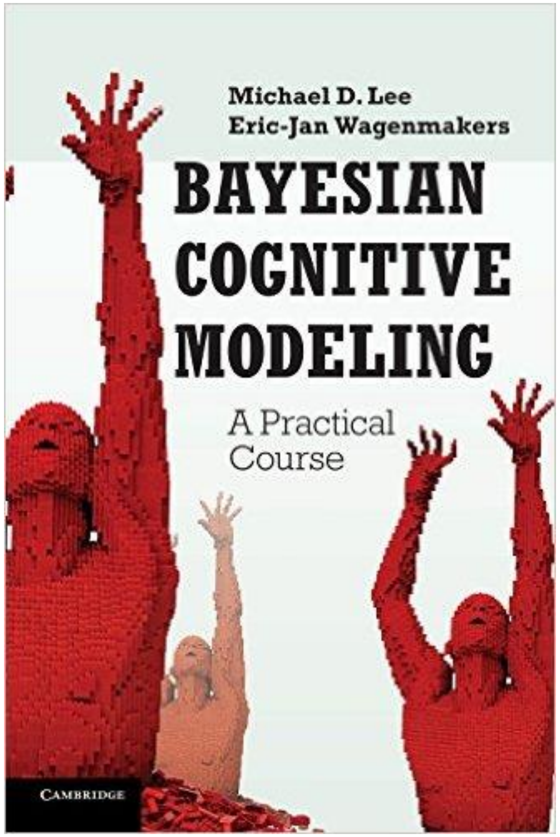
- この資料は，2015年7月18日に関西学院大学で行われた "Bayesian Cognitive Modeling: A Practical Course" (通称『コワイ本』)の読書会で使用した資料を基に加筆・訂正して作成したものです。
- 本資料は武藤が担当した第3章の要約です。二項分布に従う変数を例に，ベイズ推定の基本を概説しています。
- 内容に誤りがあればご指摘願います。



▲表紙がコワイ。

この資料について (2)

- この本ではWinBUGSというフリーソフトを使って実際にベイズ推定をしながらその内容を理解していくというスタンスを取っています。この資料で紹介しているコードは全てWinBUGSのものです（WinBUGSの詳細は第2章）。
- この本の第6章までの内容は，著者らのサイト (http://bayesmodels.com/) からpdfで閲覧することができます。また，WinBUGS (およびR, MATLAB) のコードも同サイトからダウンロードすることができます。



▲ベイズはコワクナイ。

3章のもくじ

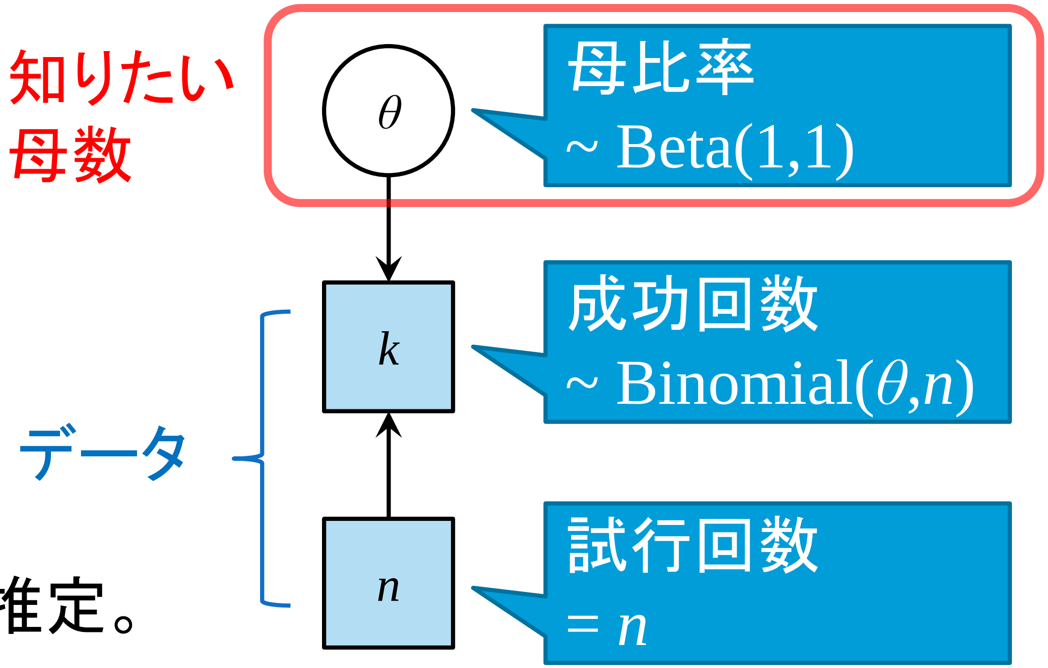
3章のもくじ

比率を推測する

（例）

コインを投げて

- 表 ＝ 成功
- 裏 ＝ 失敗　とする。



コインをn回投げてみて，

表が出る確率θの分布を推定。

ベルヌーイ試行 (Bernoulli trial)

ベルヌーイ試行

＝2値のいずれかを取る試行

- コインの表裏
- 記憶項目を正しく再生できたか
- 宝くじに当たるか当たらないか

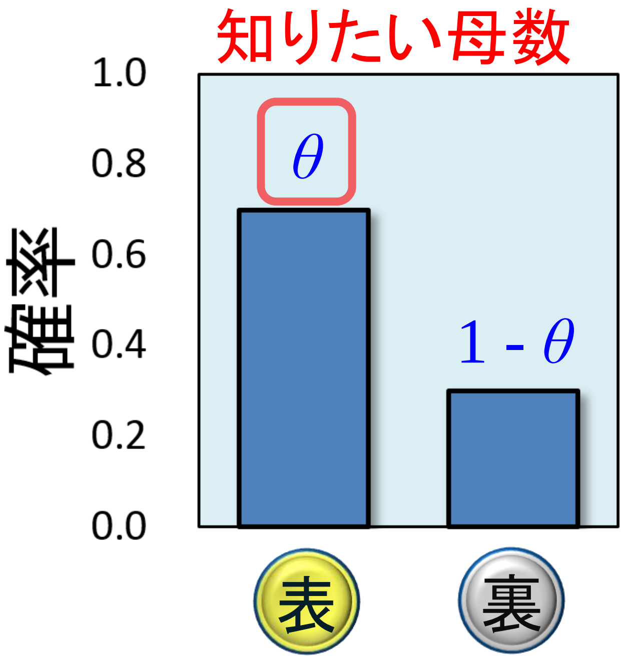


Fig. ベルヌーイ分布の例

二項分布 (binomial distribution)

独立なベルヌーイ試行 (確率θ) をn回行ったときの成功数kが従う分布。

$k \sim \mathrm{Binomial}(\theta, n)$

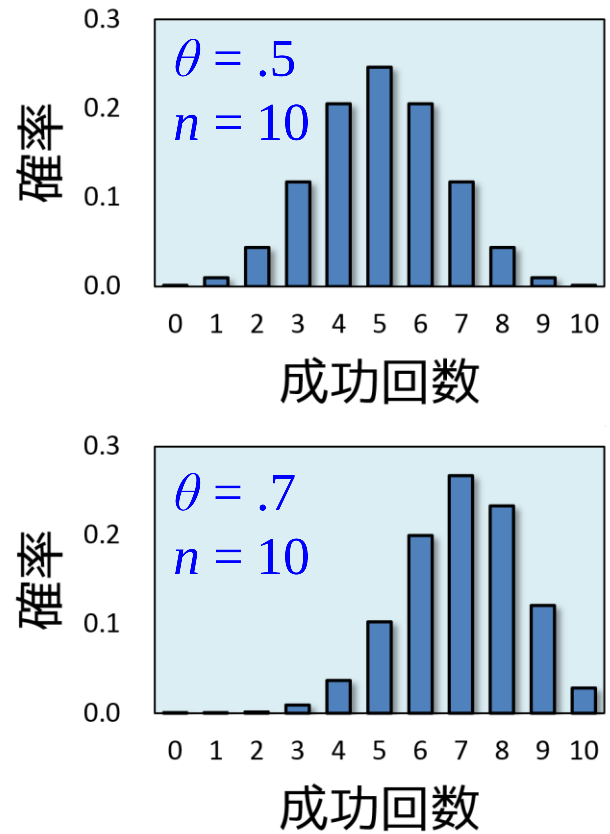


Fig. 二項分布の例

無情報事前分布 (noninformative prior distribution)

母数θについて事前の情報がない

理由不十分の原理により

無情報事前分布を設定

一様分布と同値の

ベータ分布$(\alpha = \beta = 1)$を使うと便利

$\theta \sim \mathrm{Beta}(1, 1)$

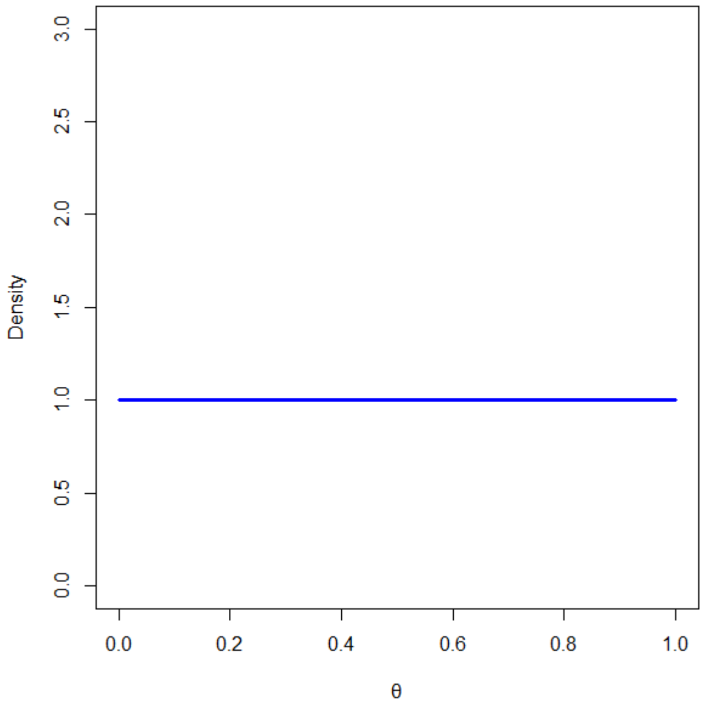
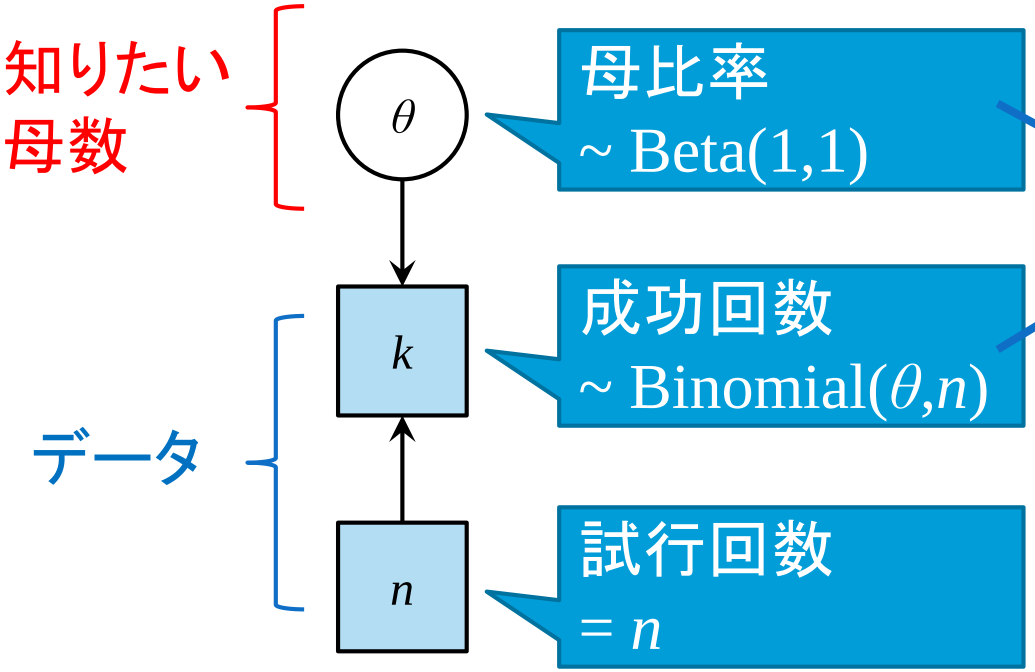


Fig. 一様分布[0,1]

モデルの確認




```
model{
theta ~ dbeta(1,1)
k ~ dbin(theta,n)
}
```

$k = 5, n = 10$ で

どうなるか？

k = 5, n = 10の事後分布

母数θについて

- 平均値 .50 (SD = .14)
- 95%信用区間[.23, .76]

→ 事前分布と比べて
中心に寄っている

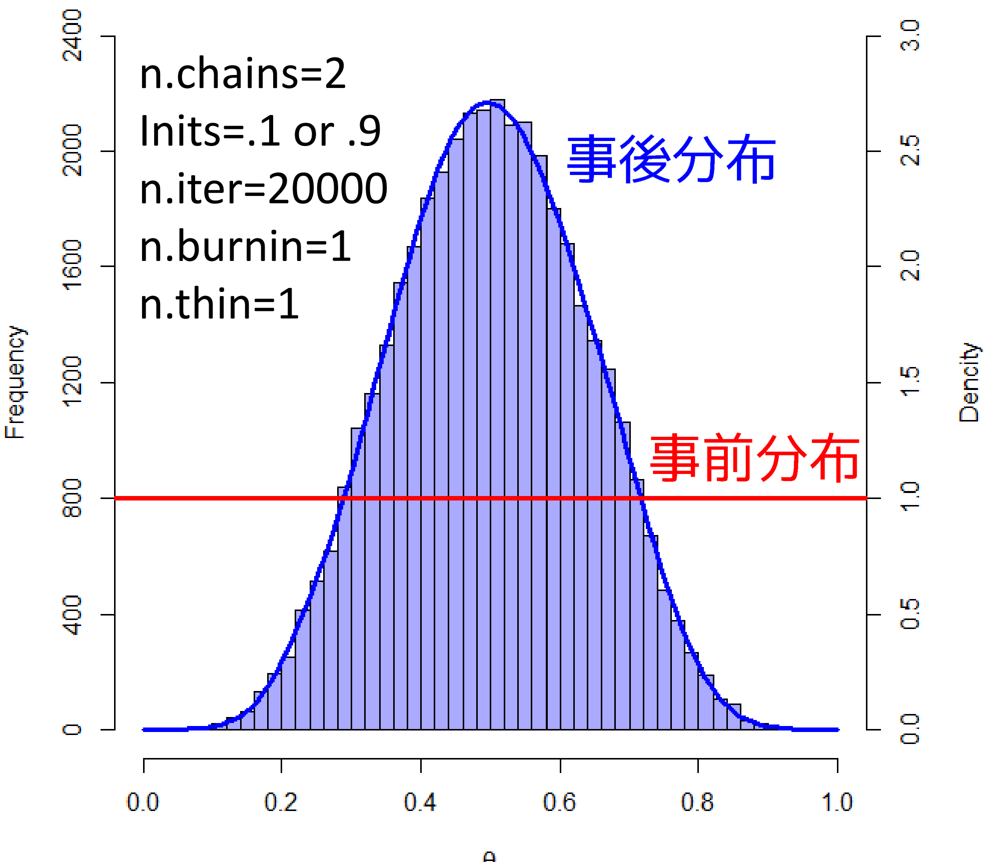


Fig. 事前分布と事後分布

要約統計量 (3.2 から先取り)

- 平均値
- 最大密度の値（モード）
- 中央値
- 95%信用区間（credible interval; 確信区間とも）
  →母数が95%の確率で含まれる区間
  （分布が左右対称なら中央の95%）
  ※分布が左右対称でない場合は選択が必要
  e.g., 最高密度区間

k = 5, n = 10の事後分布

母数θについて

- 平均値 .50 (SD = .14)
- 95%信用区間[.23, .76]

→　事前分布と比べて

　　中心に寄っている

次は

試行回数を増やしてみる。

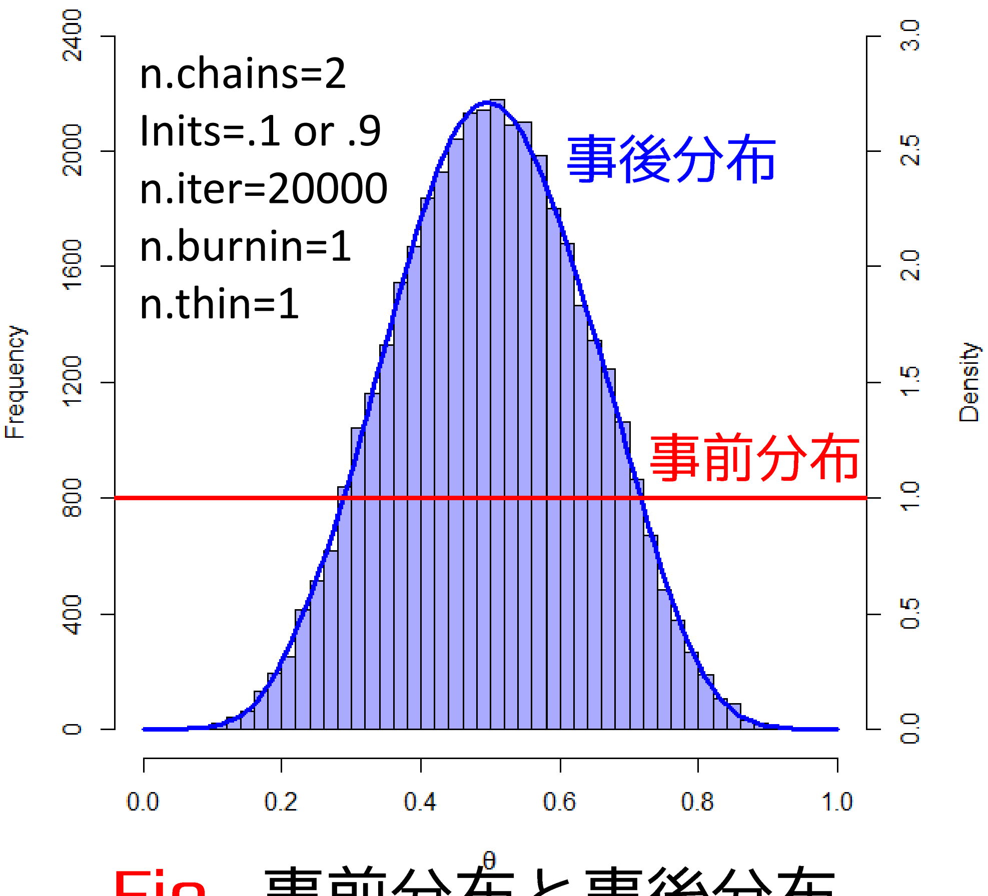


Fig. 事前分布と事後分布

k = 50, n = 100の事後分布(Ex.3)

母数θについて

- 平均値 .50 (SD = .05)
- 95%信用区間[.40, .60]

→ より中心へ

分散↓ ⇔ 精度↑

表裏の偏りは小さそう

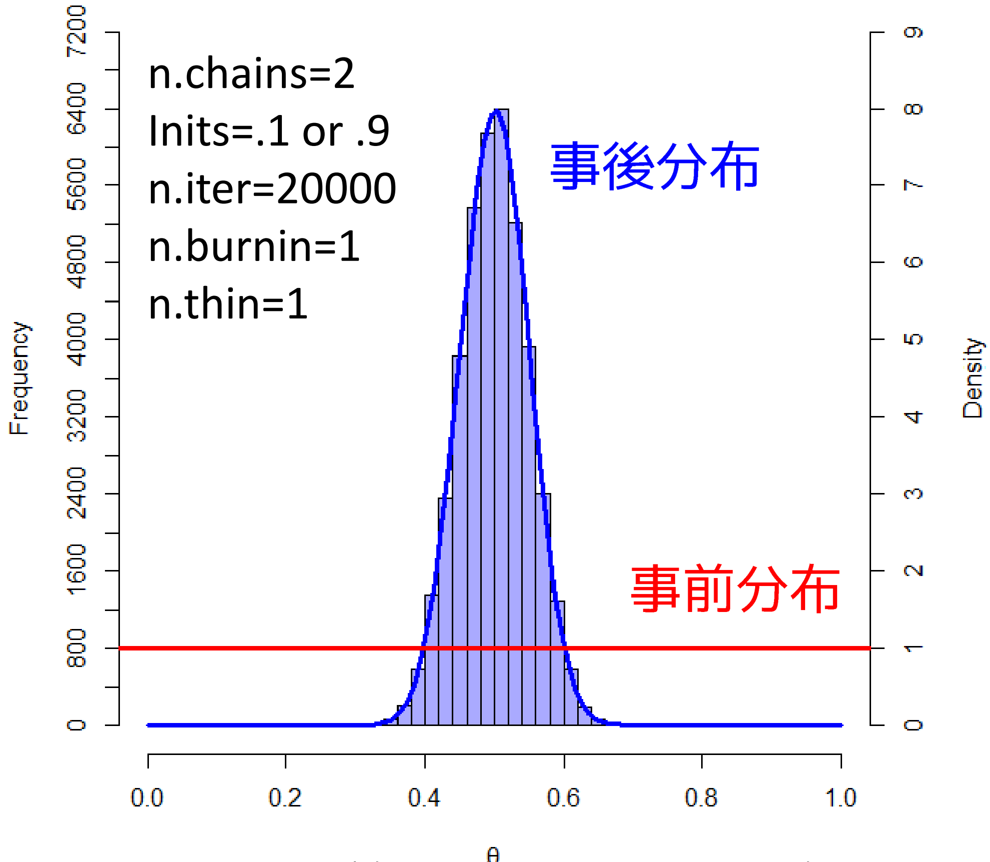
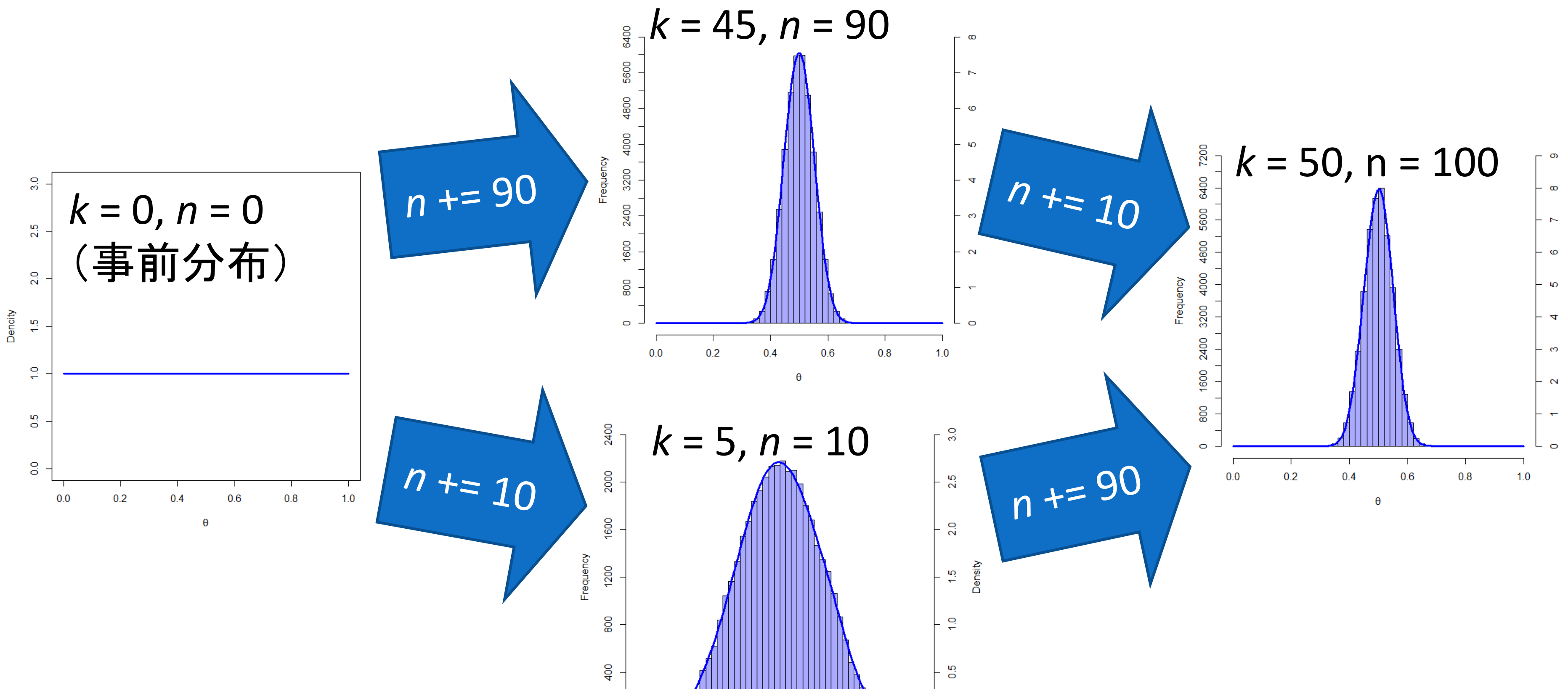


Fig. 事前分布と事後分布

追加順序の影響は無し



ベイズ更新の繰り返し

k = 99, n = 100の事後分布(Ex.5)

母数θについて

- 平均値 .98 (SD = .01)
- 95%信用区間[.95, 1.00]

表ばかり出るコイン

（磁石か何か？）

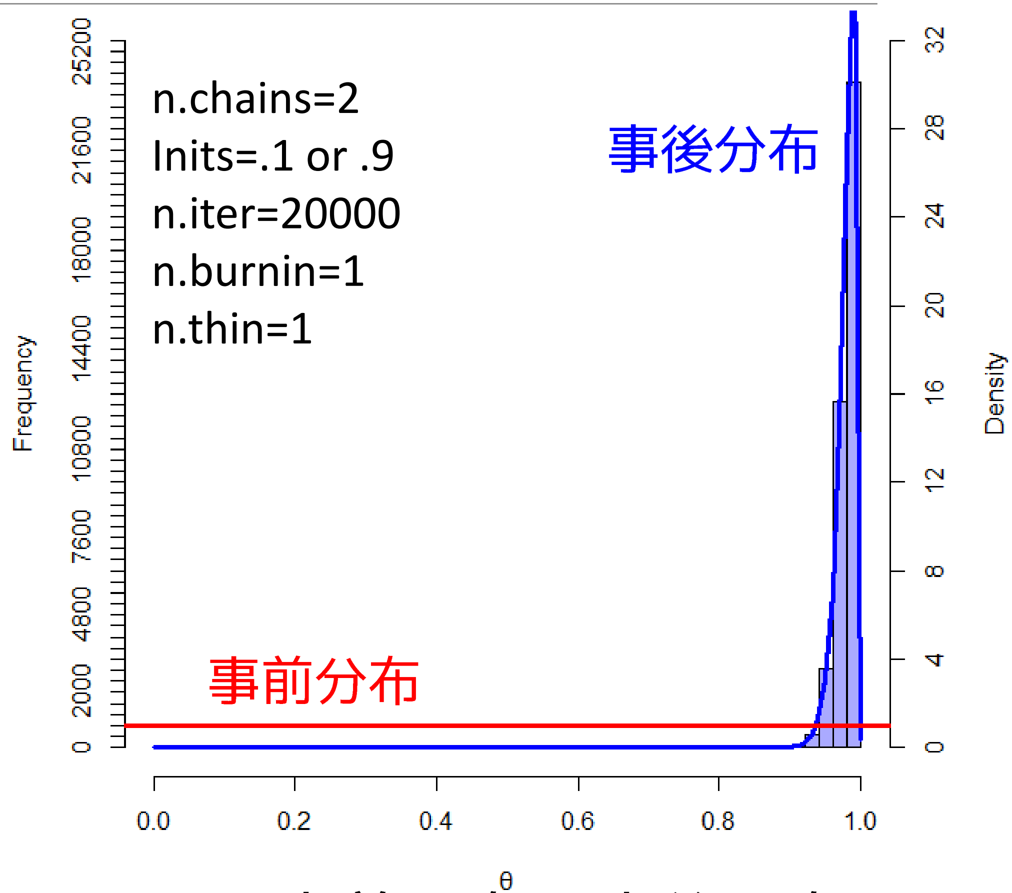


Fig. 事前分布と事後分布

k = 0, n = 1の事後分布(Ex.6)

母数θについて

- 平均値 .33 (SD = .24)
- 95%信用区間[.01, .84]

たった一回の試行では

不確かさはあまり減らない

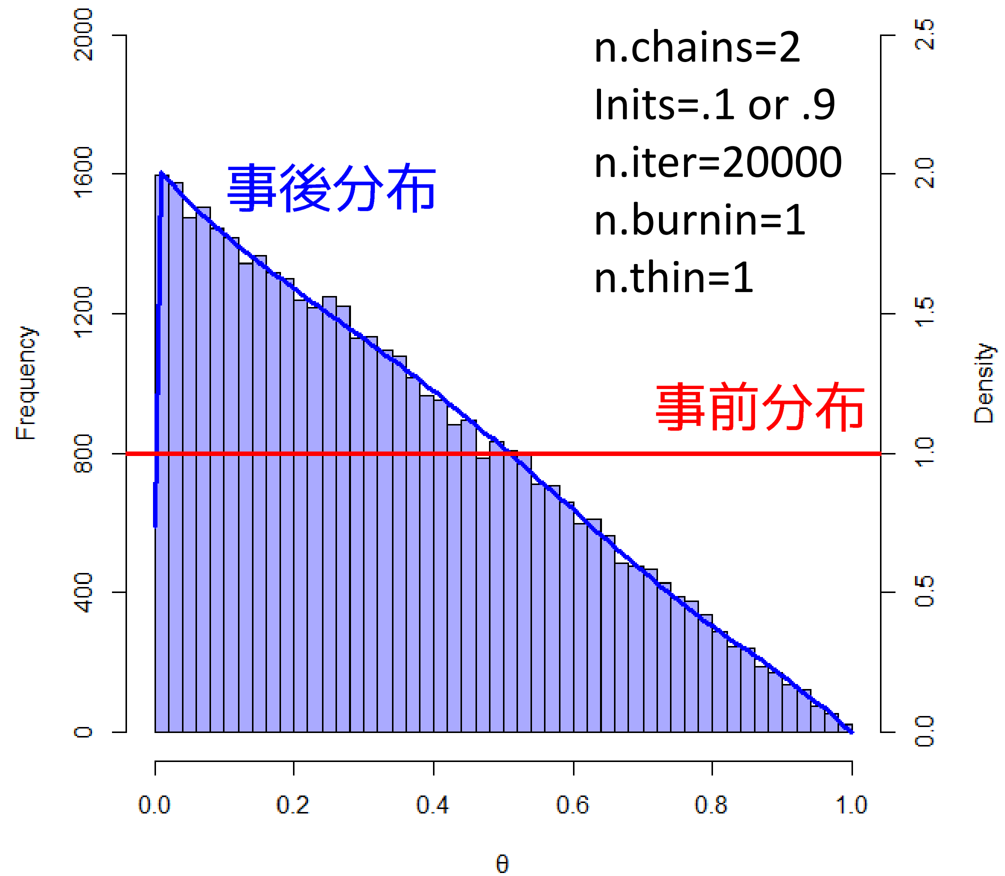


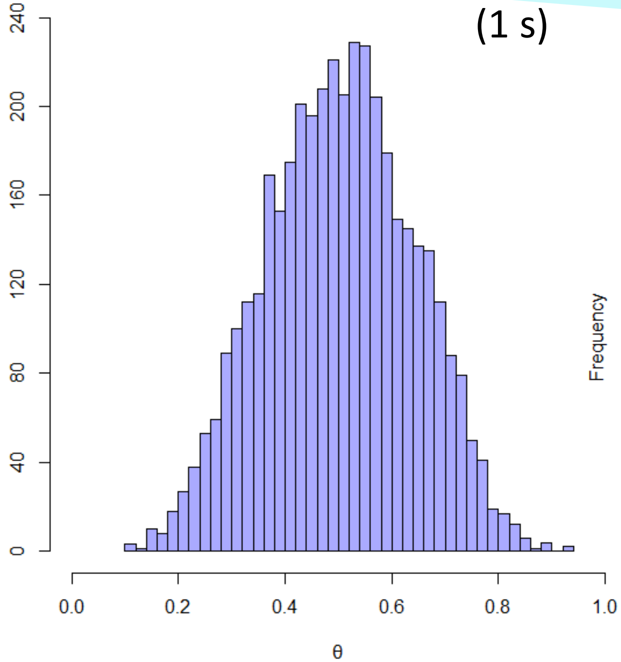
Fig. 事前分布と事後分布

サンプリング数の影響(Ex.4)

サンプリング数を増やすほど滑らかに(ただし計算時間↑)

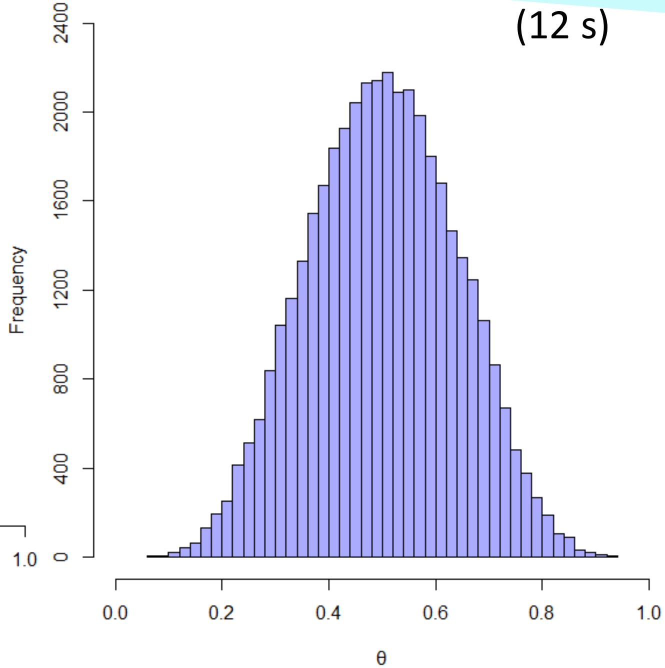
n.iter = 2,000

(1 s)



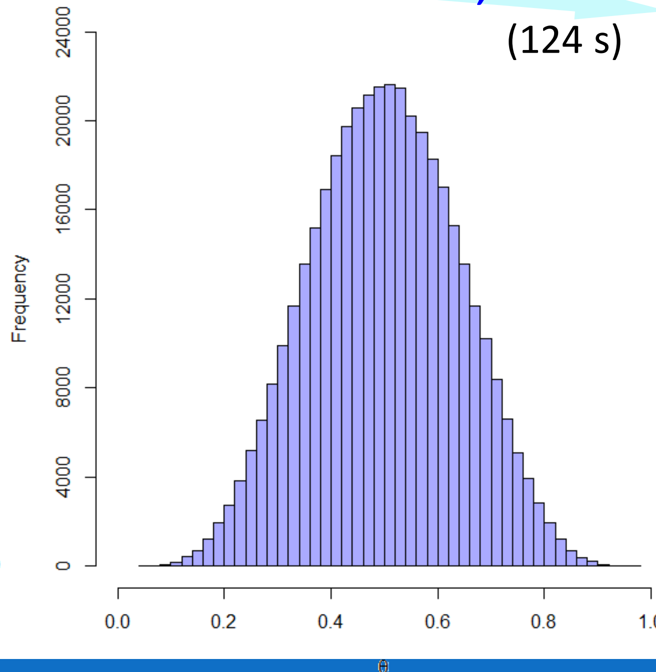
n.iter = 20,000

(12 s)



n.iter = 200,000

(124 s)



共役事前分布(conjugate prior distribution)

- 母集団分布がベルヌーイ分布のとき
  事前分布がベータ分布なら事後分布もベータ分布
- 事前と事後で分布が同じ関数になる＝共役（事前）分布
  →事後分布を解析的に求められる
- 共役分布じゃないときでもMCMCでOK

Table. 共役分布の例

母集団分布	事前分布	事後分布
ベルヌーイ分布	ベータ分布	ベータ分布
正規分布	正規分布	正規分布
正規分布	逆ガンマ分布	逆ガンマ分布
ポアソン分布	ガンマ分布	ガンマ分布

3章のもくじ

- ☑ 3.1 Inferring a rate
- ☐ 3.2 Difference between two rates
- ☐ 3.3 Inferring a common rate
- ☐ 3.4 Prior and posterior prediction
- ☐ 3.5 Posterior prediction
- ☐ 3.6 Joint distributions

比率の差の例

- コインAとコインBで表の出る確率に違いがあるか
- 統制群と実験群で〇×クイズの正答率がどう違うか
- 人間科学部と工学部ではどのくらい男女比が違うか

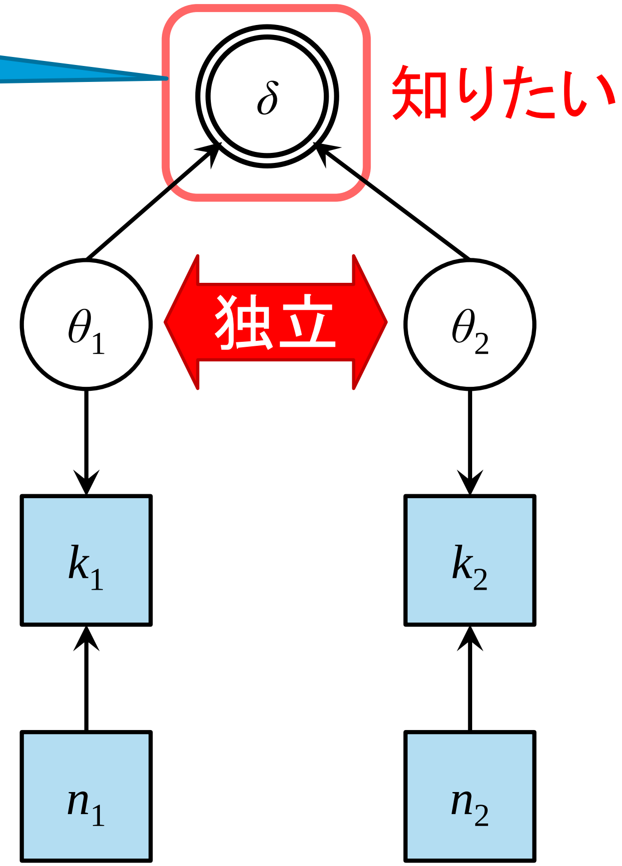
2つの比率の差を推測する

母比率の差
$= \theta_1 - \theta_2$

母比率
$\sim \mathrm{Beta}(1,1)$

成功回数
$\sim \mathrm{Binomial}(\theta_i, n)$

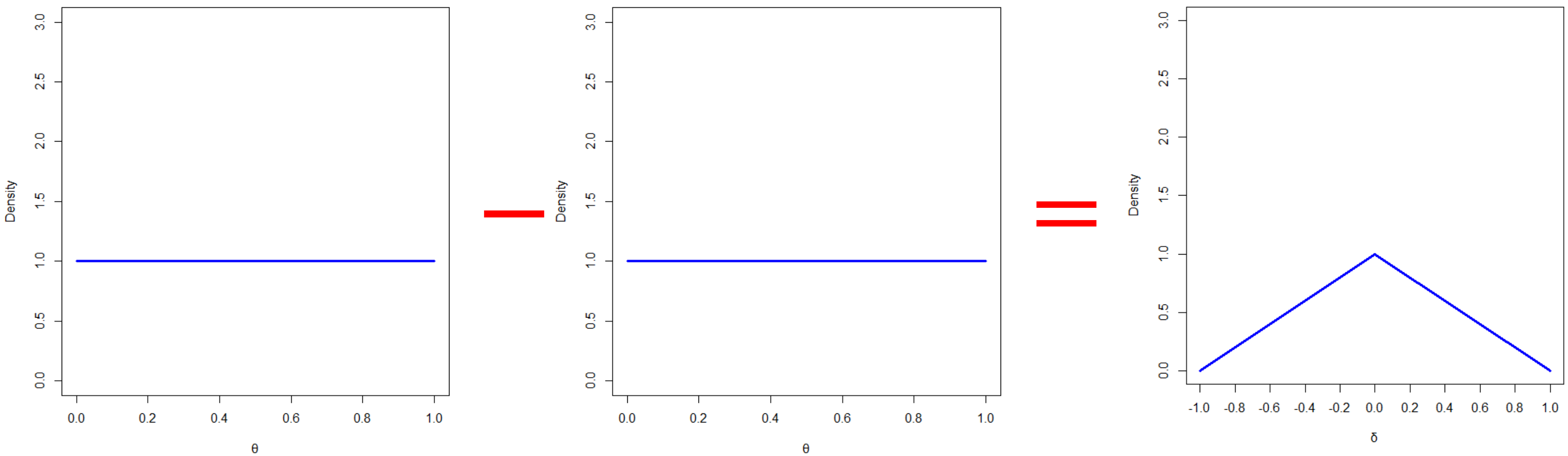
試行回数
$= n_i$



$i = \{1, 2\}$

注：一様分布の差

互いに独立な一様分布の和（差）は一様分布ではない。



θ_1の事前分布　　θ_2の事前分布　　δ の事前分布

モデルの確認

model{

Delta <- theta1 – theta2

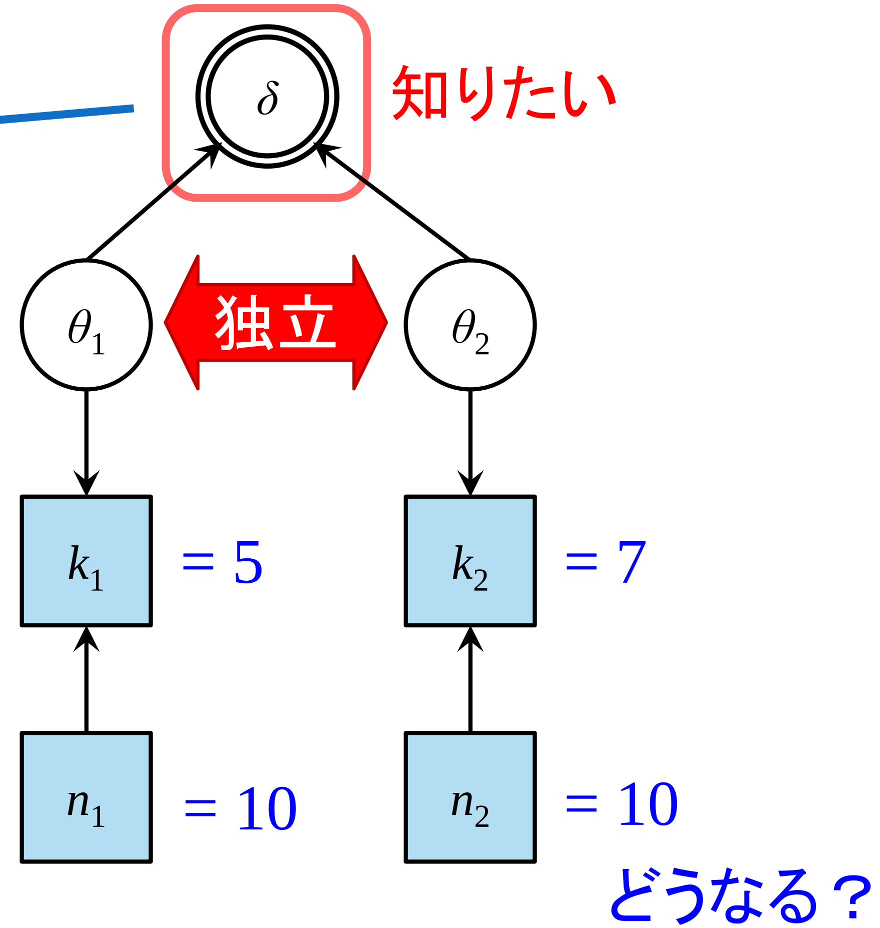
theta1 ~ dbeta(1,1)

theta2 ~ dbeta(1,1)

k1 ~ dbin(theta1,n1)

k2 ~ dbin(theta2,n2)

}



事後分布(k/n={5/10, 7/10})

母数θについて

- 平均値 -.17 (*SD* = .19)
- 95%信用区間[-.52, .21]

→　ややマイナス寄り

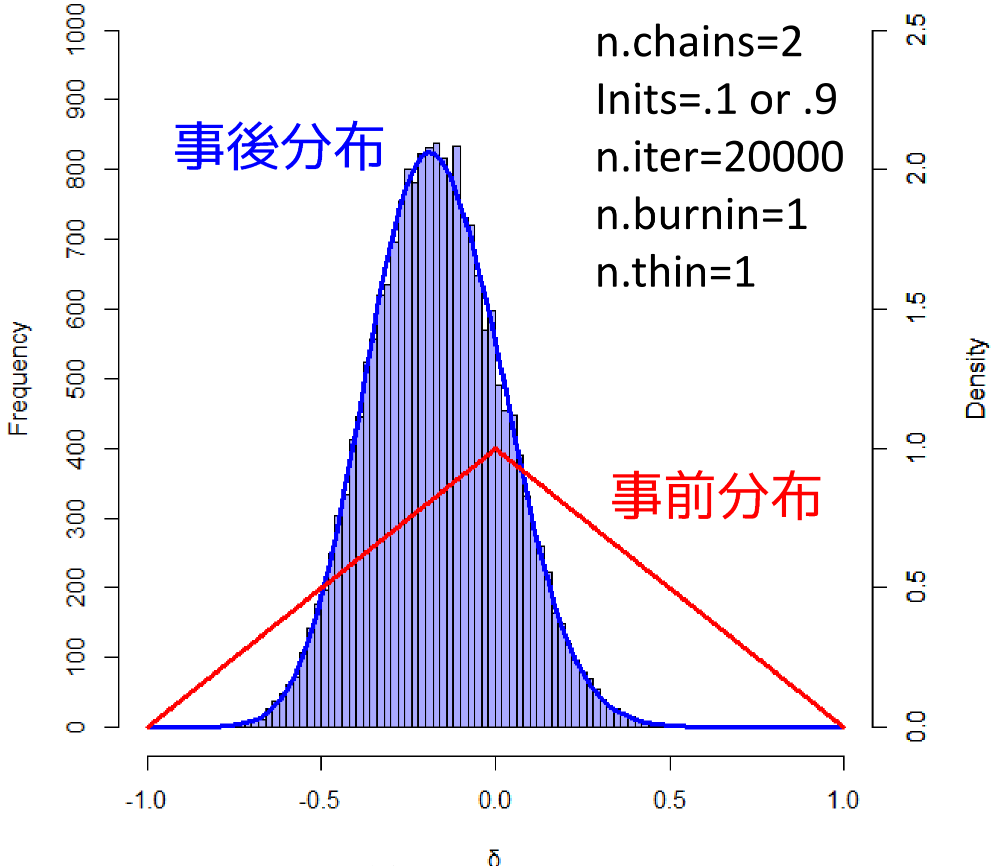
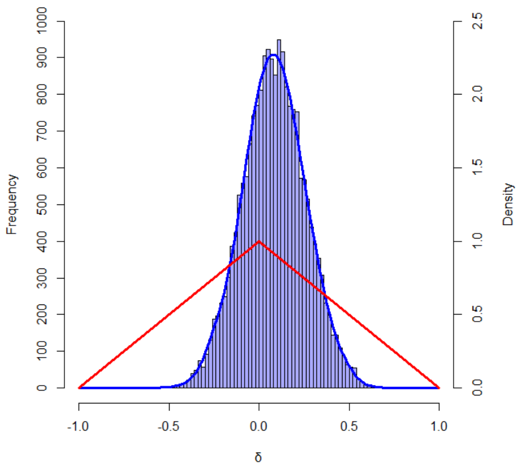
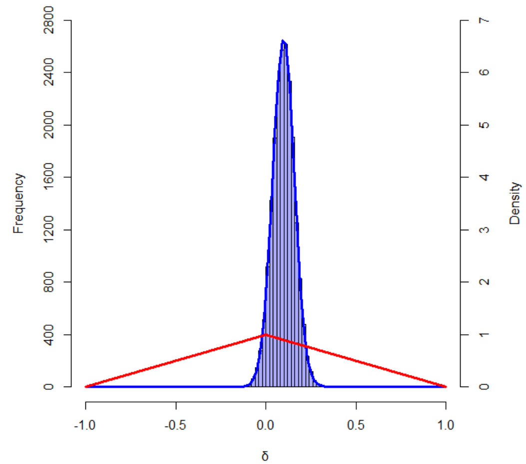


Fig. 事前分布と事後分布

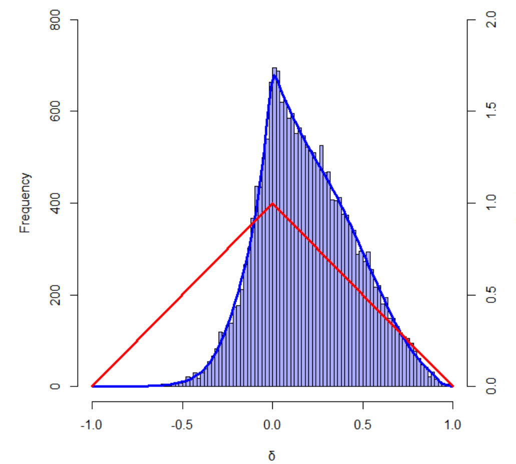
いろいろなデータの事後分布



$k_1 = 8, n_1 = 10,$
$k_2 = 7, n_2 = 10$



$k_1 = 80, n_1 = 100,$
$k_2 = 70, n_2 = 100$



$k_1 = 0, n_1 = 1,$
$k_2 = 0, n_2 = 5$

(Ex. 1)

(Ex. 2)

3章のもくじ

- ☑ 3.1 Inferring a rate
- ☑ 3.2 Difference between two rates
- ☐ 3.3 Inferring a common rate
- ☐ 3.4 Prior and posterior prediction
- ☐ 3.5 Posterior prediction
- ☐ 3.6 Joint distributions

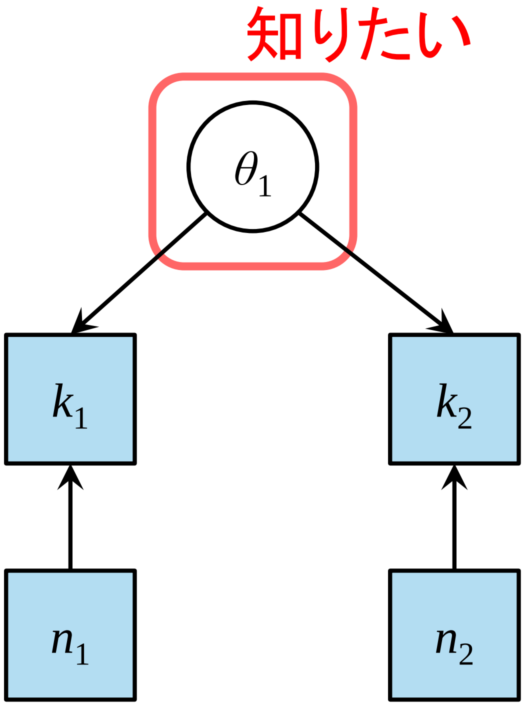
共通の比率を推測する

母比率
~ Beta(1,1)

成功回数
~ Binomial(θ_i,n)

試行回数
= n_i

i = {1, 2}



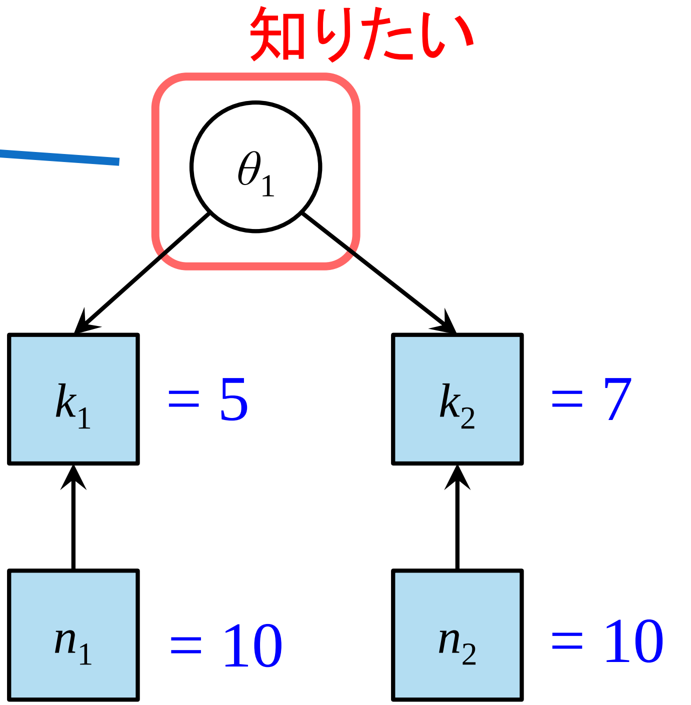
共通の比率の例

- コイントスを一日目に10回，二日目に20回やってみて，表が出る確率を調べる。
- 東京で100人，大阪で70人に〇×クイズを実施して，日本人の正解率を推測する。

モデルの確認

```
model{
theta ~ dbeta(1,1)

k1 ~ dbin(theta,n1)
k2 ~ dbin(theta,n2)
}
```



どうなる？

事後分布(k/n={5/10, 7/10})

母数θについて

- 平均値 .59 (SD = .10)
- 95%信用区間[.38, .78]

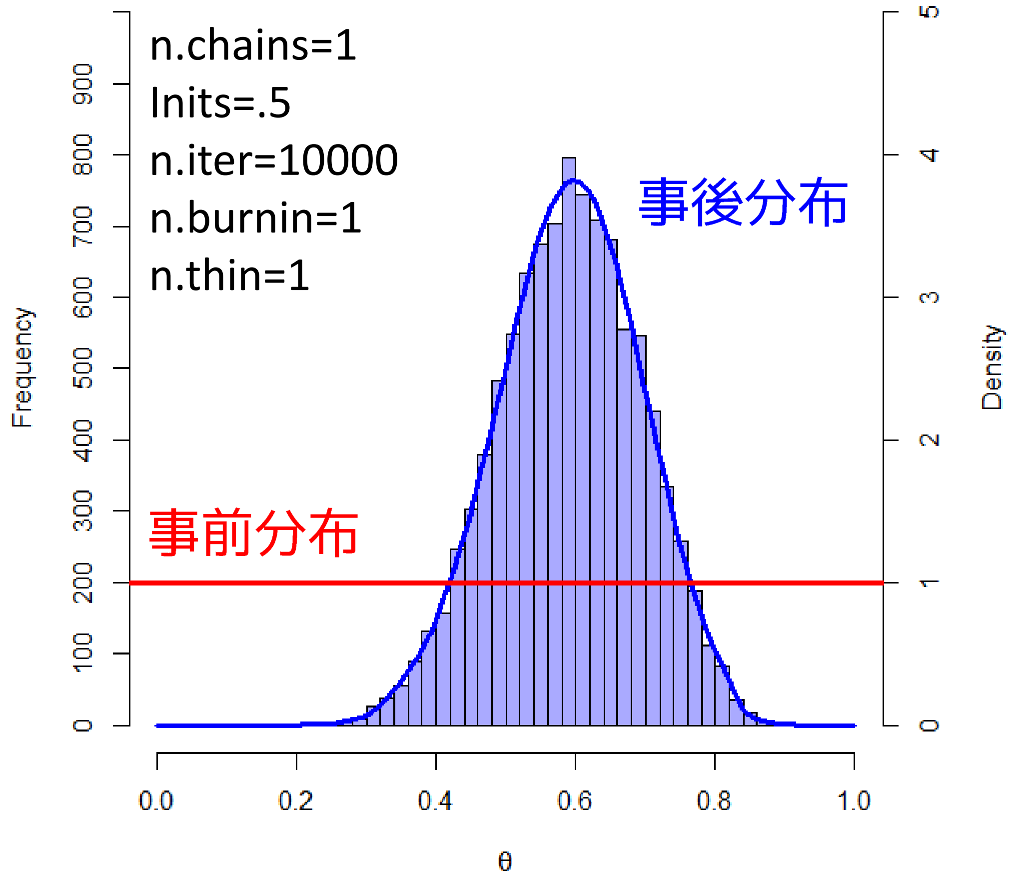


Fig. 事前分布と事後分布

k/n={14/20, 16/20} のとき (Ex. 1)

母数θについて

- 平均値 .74 (SD = .07)
- 95%信用区間[.60, .86]

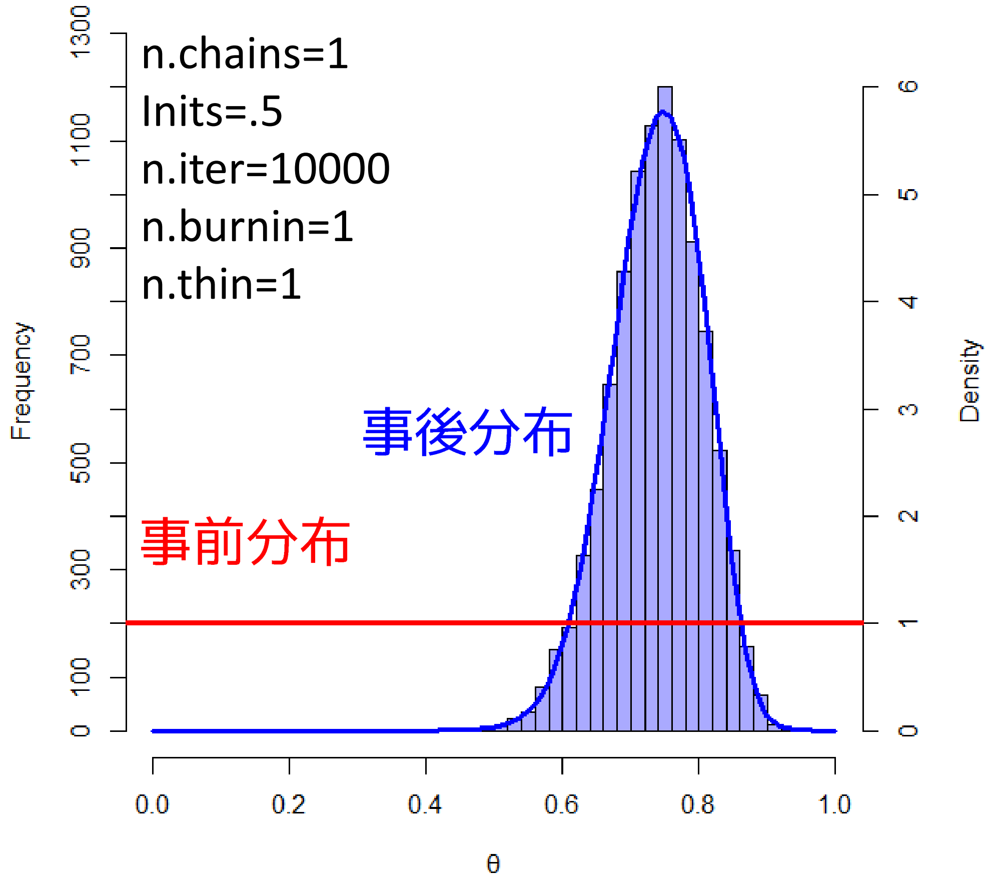


Fig. 事前分布と事後分布

サンプルが均質でないとき (Ex. 2)

$k_1 = 0$, $n_1 = 10$,
$k_2 = 10$, $n_2 = 10$

母数θについて

- 平均値 .50 (SD = .10)
- 95%信用区間[.30, .70]

→綺麗な左右対称のグラフ

しかし……

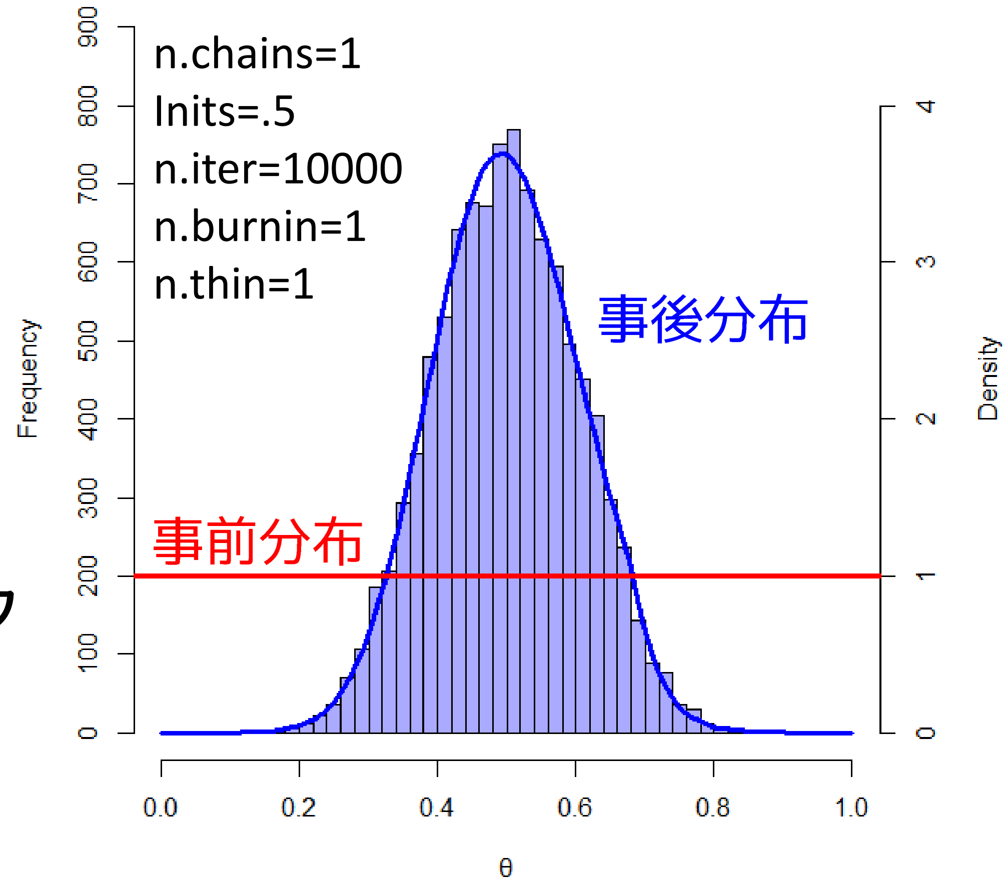


Fig. 事前分布と事後分布

不適切なモデル？

サンプルが均質でない

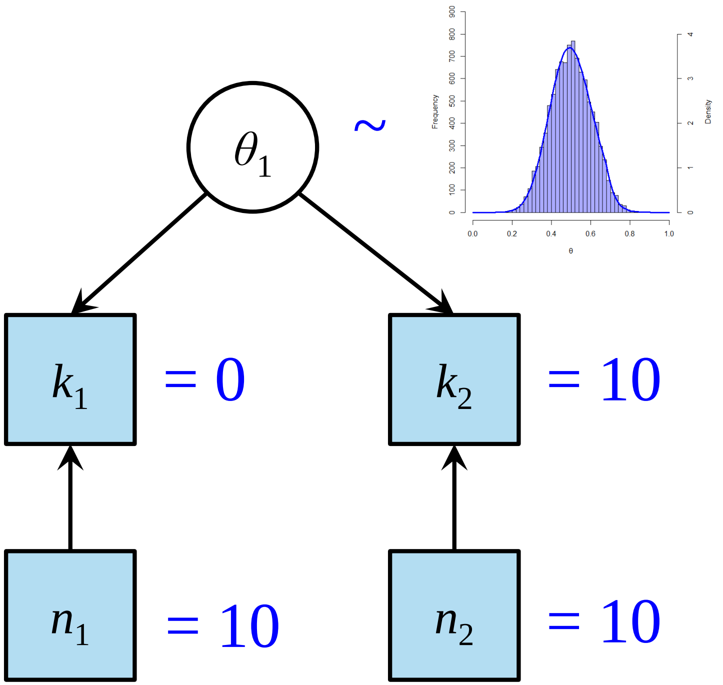
共通の比率を持つという

前提が怪しい

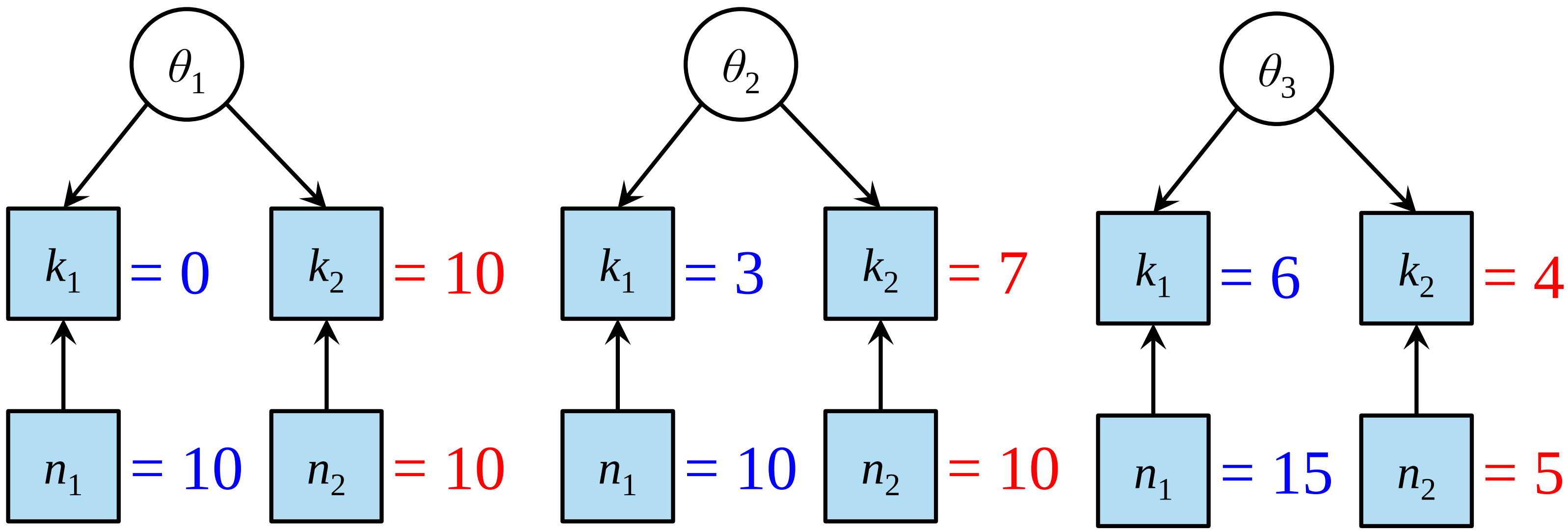
このモデルは不適切？

先の差のモデルのほうがベター？

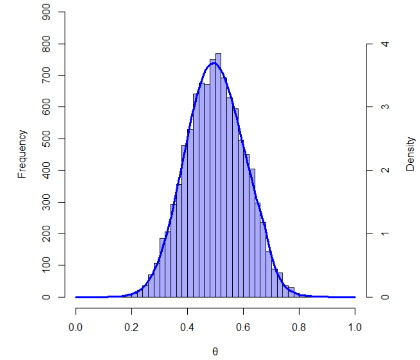
研究の関心に合わせてモデルを選択



同一の事後分布 (Ex. 3)



どのモデルでも全く同じ事後分布が推定される



∵最終的に $k_{\mathrm{total}} = 10,\ n_{\mathrm{total}} = 20$ となるから

＝情報を追加する順序は無関係

3章のもくじ

- ☑ 3.1 Inferring a rate
- ☑ 3.2 Difference between two rates
- ☑ 3.3 Inferring a common rate
- ☐ 3.4 Prior and posterior prediction
- ☐ 3.5 Posterior prediction
- ☐ 3.6 Joint distributions

事前予測と事後予測

ベイズ分析では4種類の分布が利用できる。

- 事前分布　prior distribution
- 事前予測分布　prior predictive distribution（＝周辺尤度）

  →事前分布から,

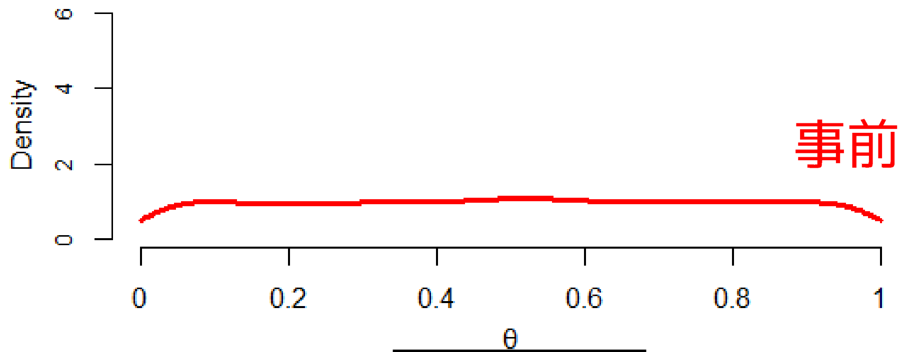
  どんな観測データが得られると予測できるか？
- 事後分布　posterior distribution
- 事後予測分布　posterior predictive distribution

  →事後分布から,

  次にどんな観測データが得られると予測できるか？

事前分布と事前予測分布 (Ex. 1)

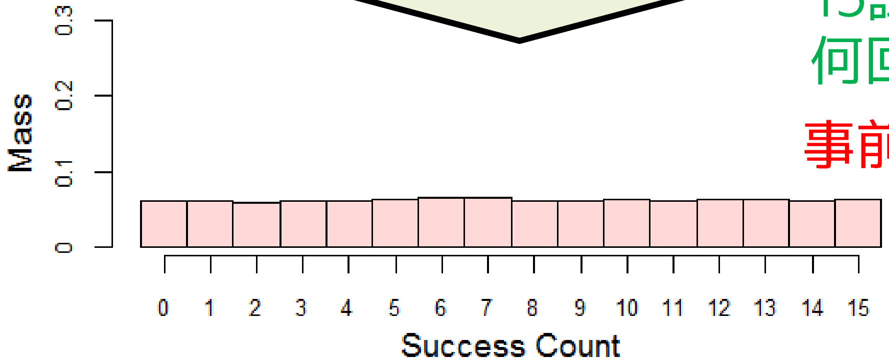
母数θの分布



事前分布

観測データの予測分布

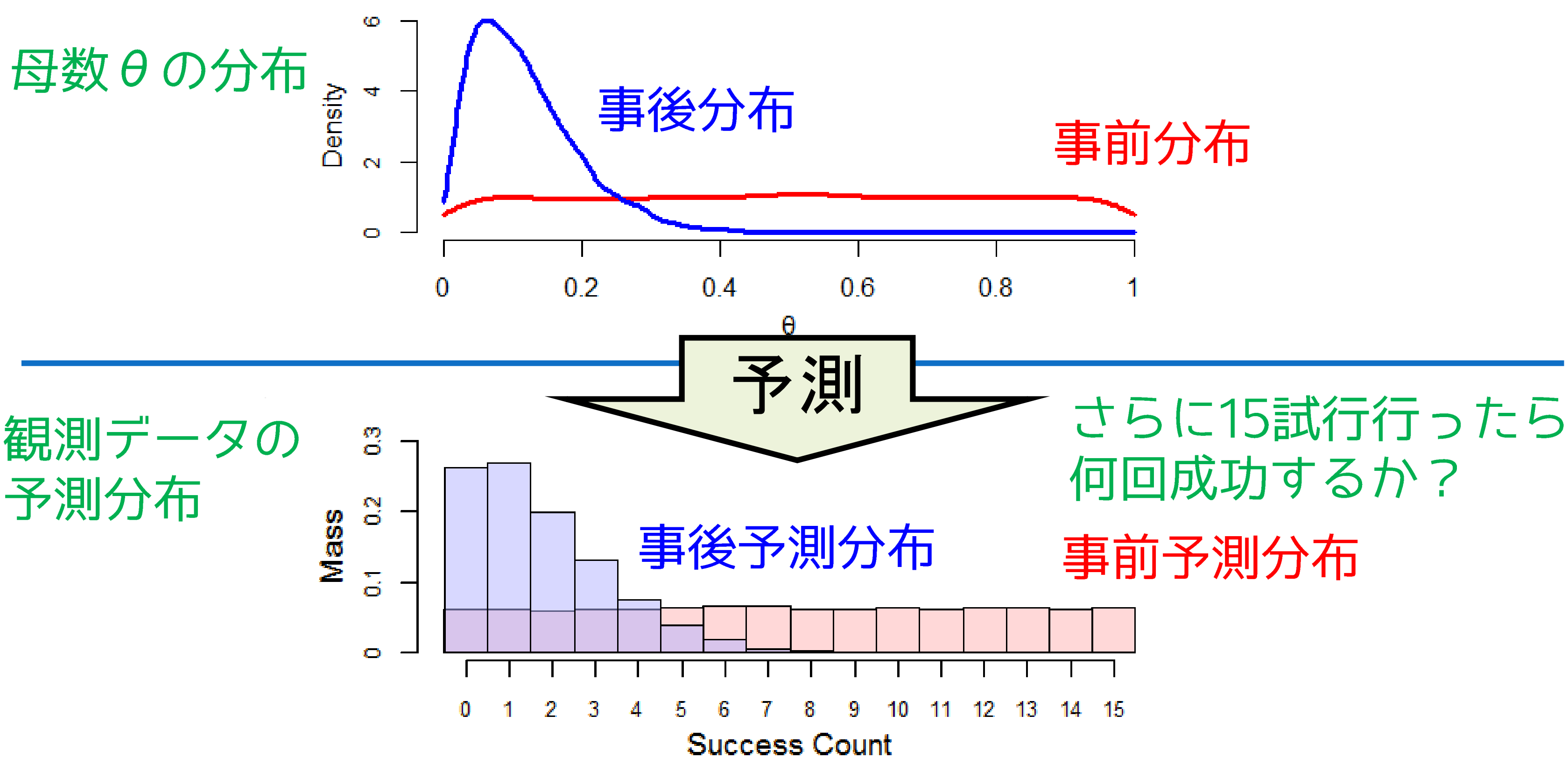
15試行行ったら
何回成功するか？



事前予測分布

事後分布と事後予測分布 (Ex. 1)

実際にデータを取ったら15回中1回だけ成功だった。

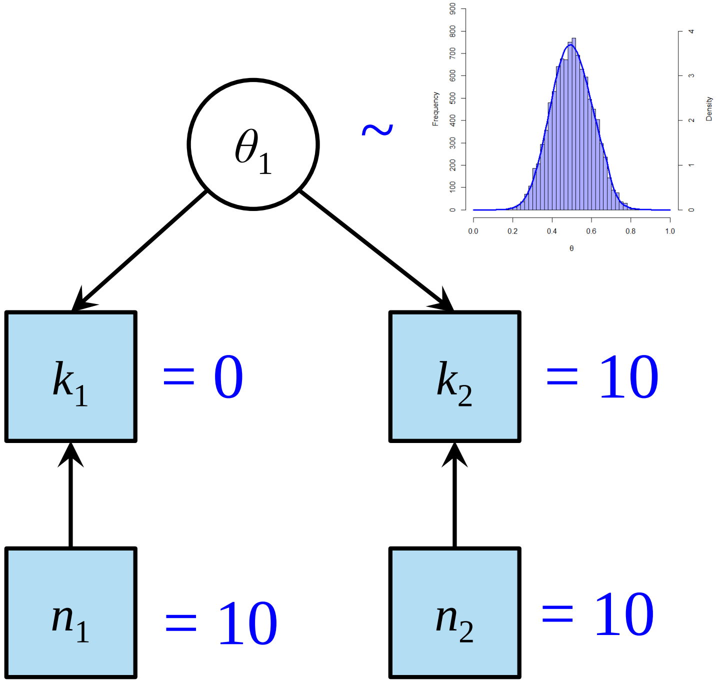


3章のもくじ

- ☑ 3.1 Inferring a rate
- ☑ 3.2 Difference between two rates
- ☑ 3.3 Inferring a common rate
- ☑ 3.4 Prior and posterior prediction
- ☐ 3.5 Posterior prediction
- ☐ 3.6 Joint distributions

事後予測分布の使用例

- 事後予測分布からモデルの適切さ (adequacy) を評価



3.3節で挙げた例の再考：

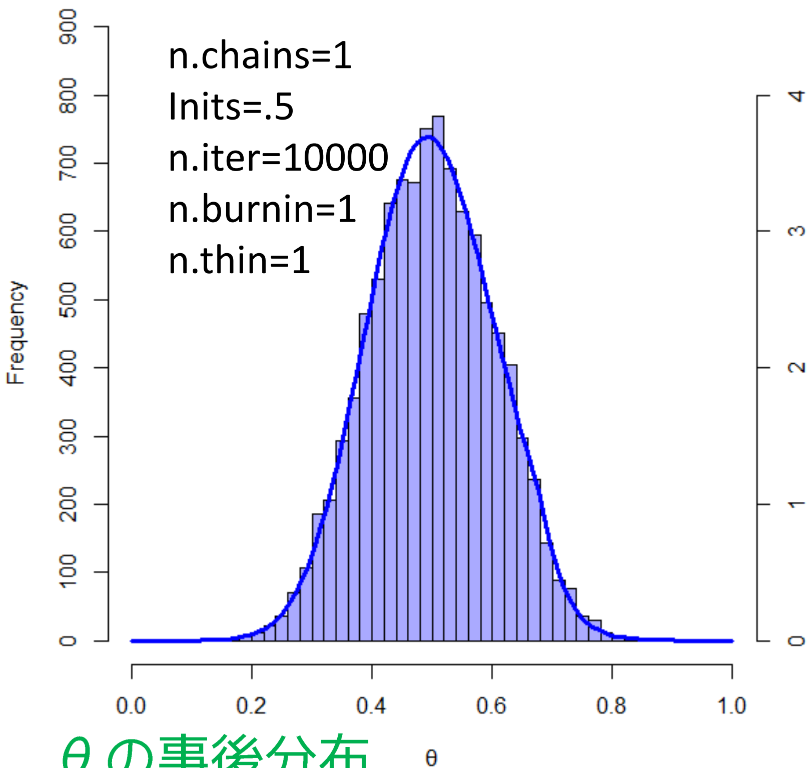
共通のθ

→ k_1 と k_2 は相関するはず

推定された事後分布から

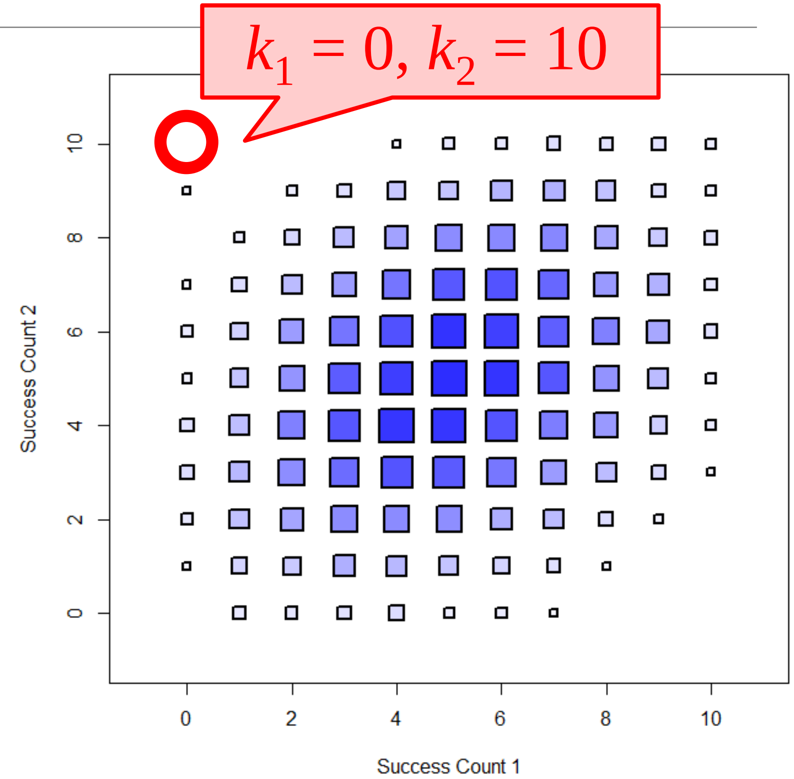
元のデータを予測できるか？

事後分布と事後予測分布



θ の事後分布

- 平均値 .50 (SD = .10)
- 95%信用区間[.30, .70]



k_1とk_2の事後予測分布

（大きくて濃い＝起こりやすい）

∴このモデルは適切とはいいがたい

3章のもくじ

- ☑ 3.1 Inferring a rate
- ☑ 3.2 Difference between two rates
- ☑ 3.3 Inferring a common rate
- ☑ 3.4 Prior and posterior prediction
- ☑ 3.5 Posterior prediction
- ☐ 3.6 Joint distributions

同時分布の推測

- これまでは単一の分布のみを推測してきた。

  e.g., 与えられたnとkからθの分布を推測

- 複数の変数に関心があり，かつそれらが互いに影響を及ぼす場合は？

  e.g., kのみが与えられたときにnとθの分布はどうなるか

  →同時分布 (joint distribution) を推測する

同時分布の具体例

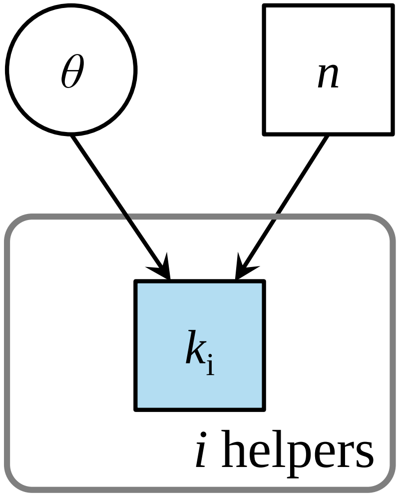
- 5人の助手がアンケートを実施
- どの助手もn部ずつ配布 (n_{max} = 500)
- 調査対象者は無作為抽出
- 返ってきたアンケートの部数は k = {16, 18, 22, 25, 27}
- 1人の助手が配った部数nは？
- 回収率θは？

データの整理とモデル

	配布部数	回収部数	回収率
助手A	n	16	θ
助手B	n	18	θ
助手C	n	22	θ
助手D	n	25	θ
助手E	n	27	θ

母数　　データ　　母数

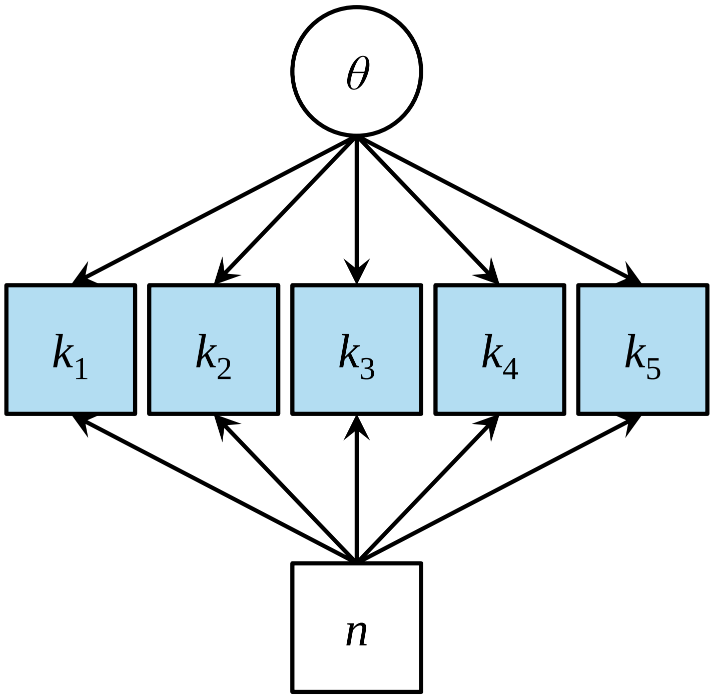
$(n_{\mathrm{max}} = 500)$



データの整理とモデル

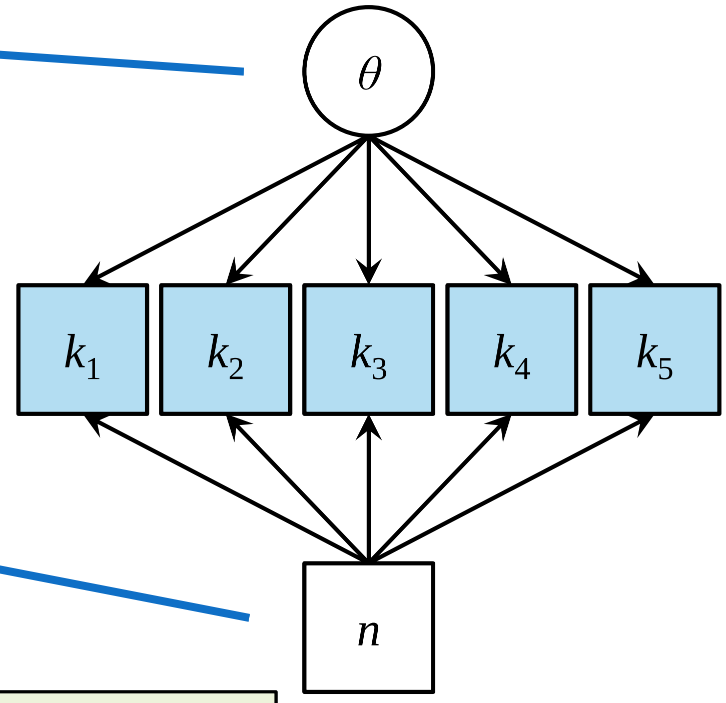
	配布部数	回収部数	回収率
助手A	n	16	θ
助手B	n	18	θ
助手C	n	22	θ
助手D	n	25	θ
助手E	n	27	θ
	母数	データ	母数

(n_{max} = 500)



モデルの確認

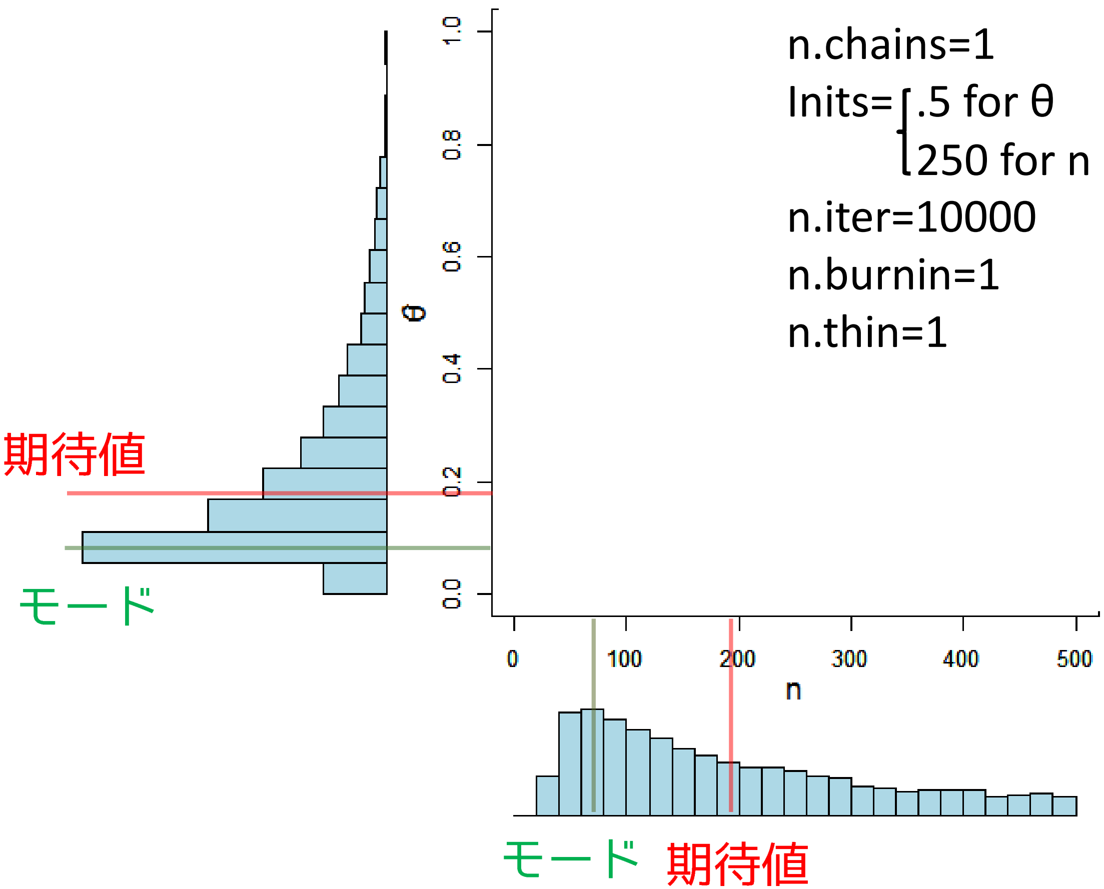
```
model{
theta ~ dbeta(1,1)
for (i in 1:m){
   k[i] ~ dbin(theta, n)
}
n ~ dcat(p[])
for (i in 1:nmax){
   p[i] ~ <- 1/nmax
}
}
```



1〜500の整数が

等確率で出現する事前分布

2つの母数の事後分布



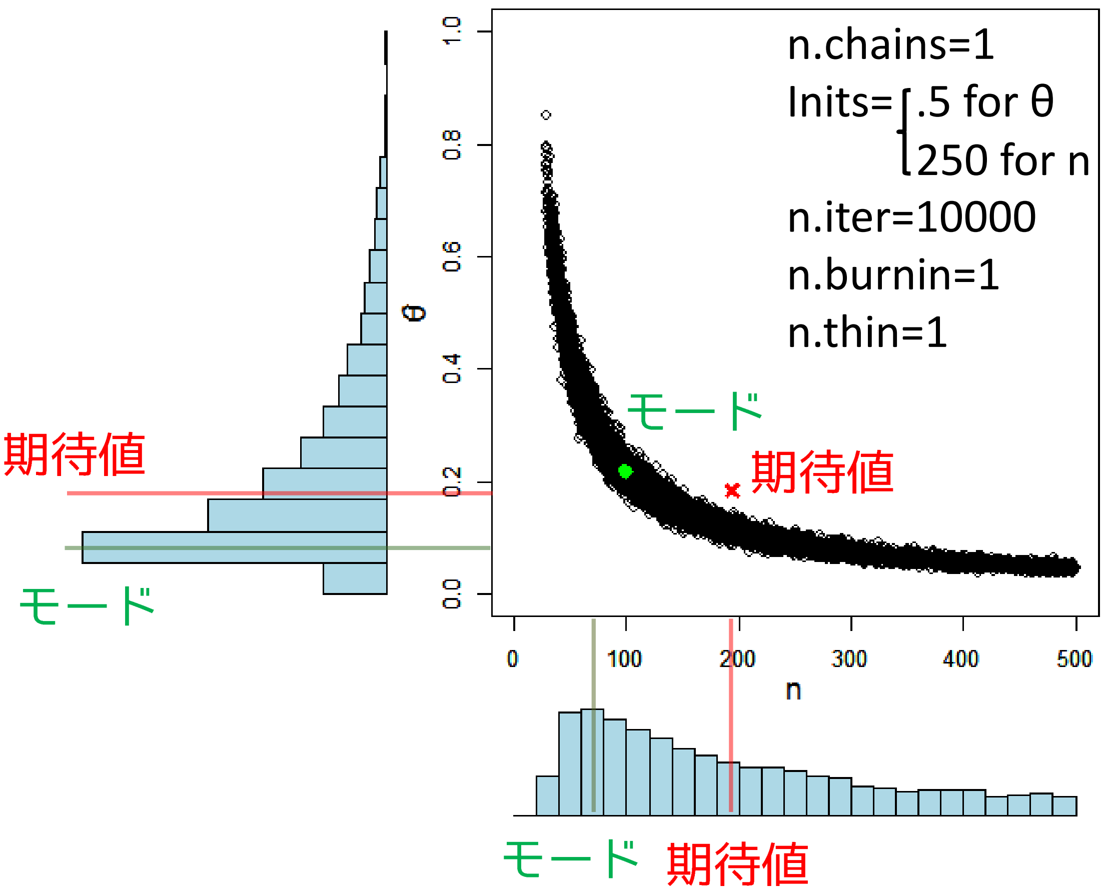
- θの期待値 = .183
- nの期待値 = 194

→これだけの情報から

結論を導くのは危険

(Ex. 1)

2つの母数の同時事後分布



同時分布を考慮

→θとnはトレードオフ

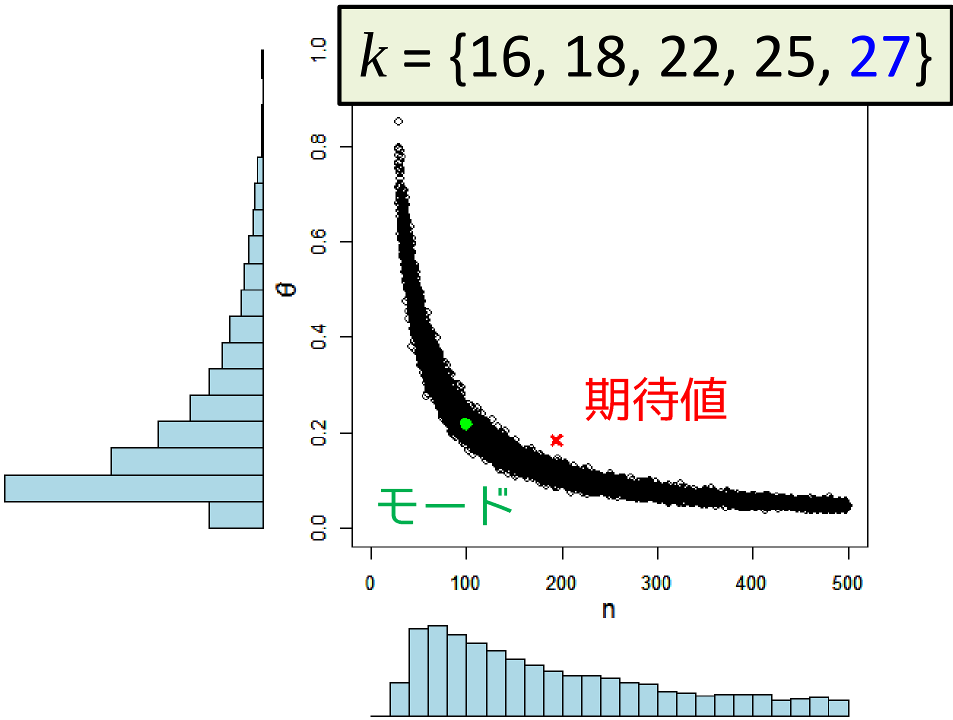
（しかも非線形）

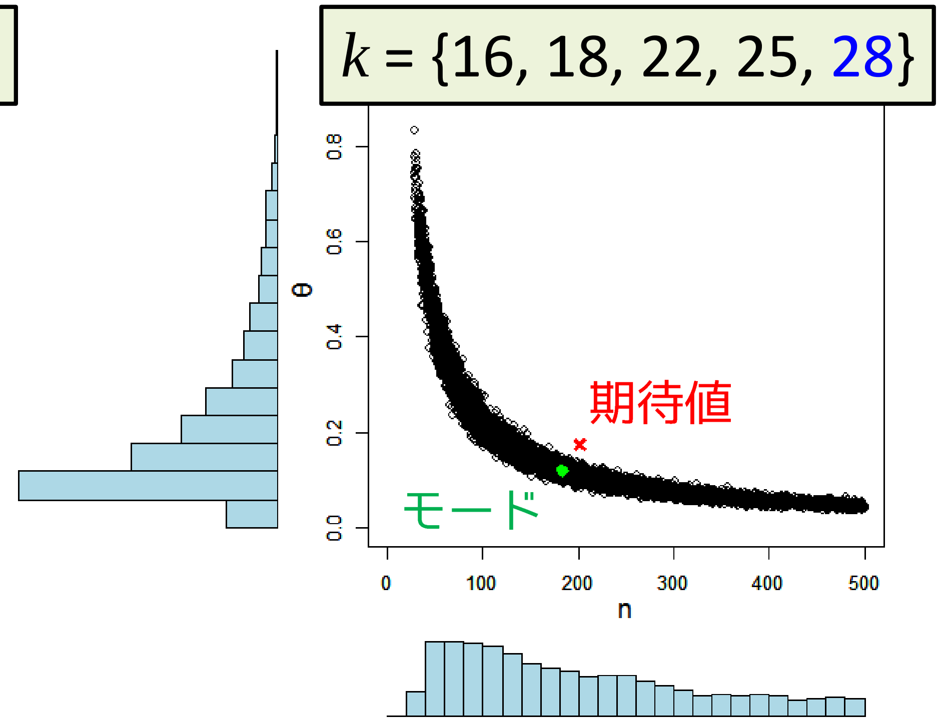
代表値としては

同時分布のモードが

より適切 (Ex. 2)

ちょっとだけデータを変える

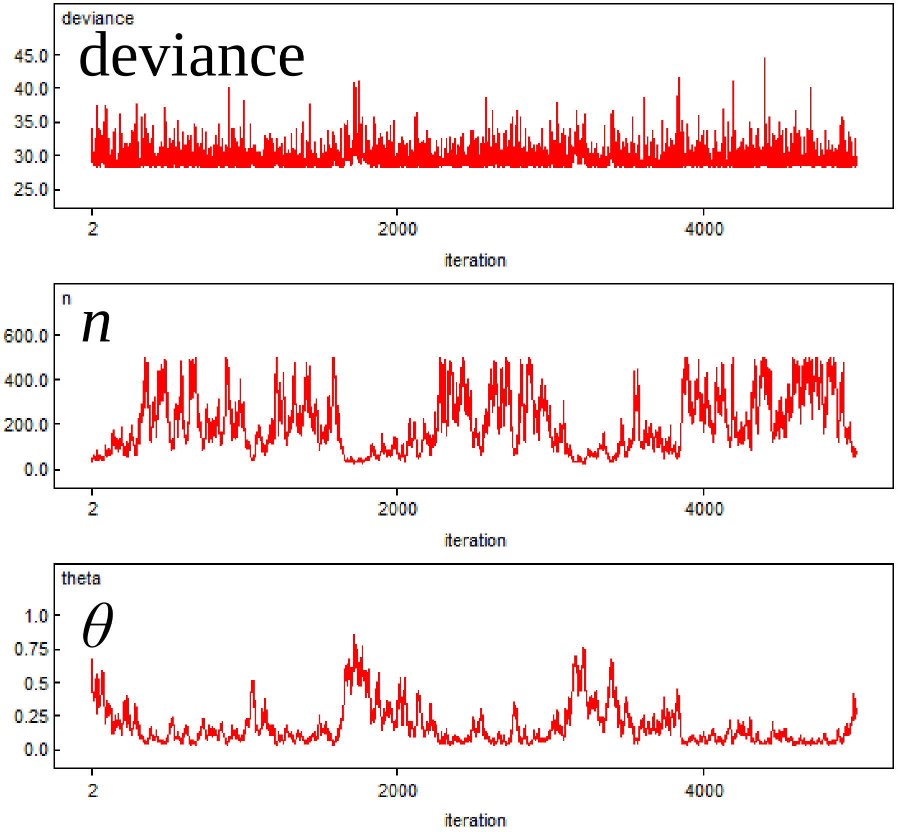




モードの位置が大きく変化 (Ex. 3)

→そもそも情報が少ないことに起因

トレースプロットを見る (Ex.4)



規則的に上下

＝自己相関を持つ

(autocorrelated)

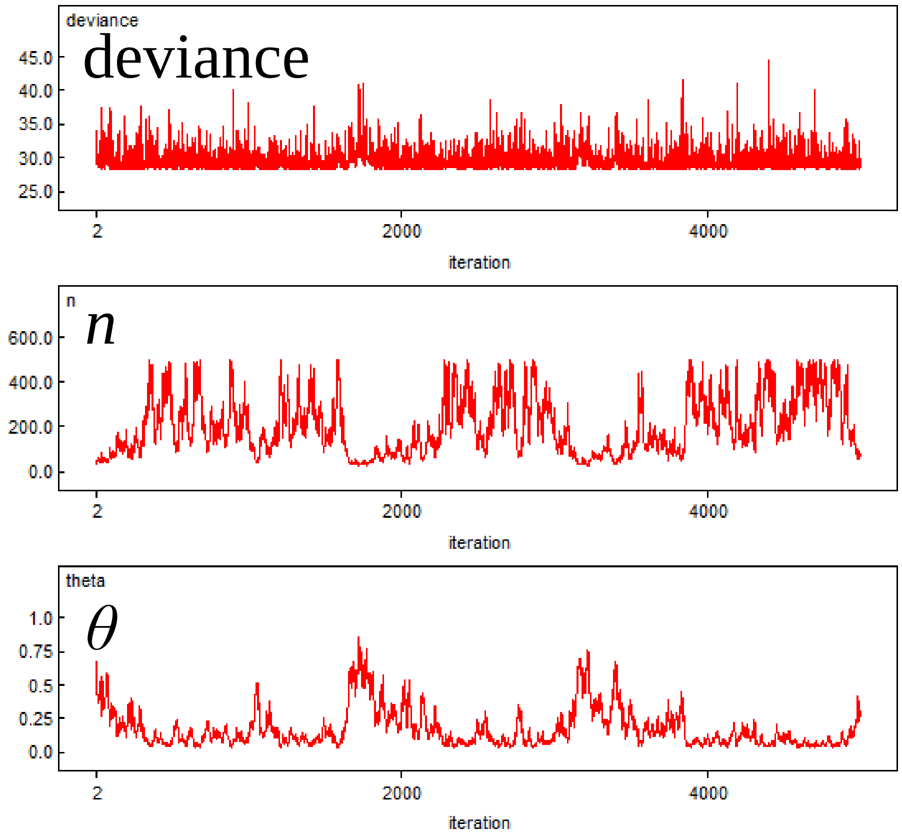
計算の効率が悪くなるため

あまり望ましくない。

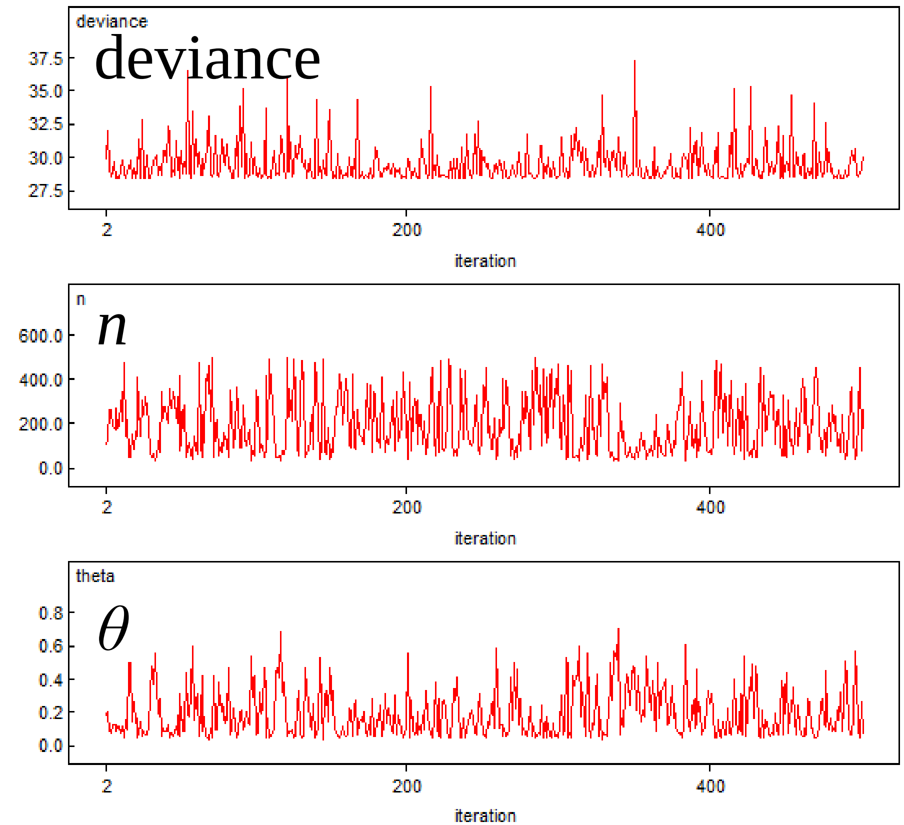
thinを1→100にしてみる

thin = 100の場合と比較 (Ex.4)

thin = 1, n.iter = 5,000



thin = 100, n.iter = 50,000



この章のまとめ

- 二項分布 (ベルヌーイ分布) に従う変数について，いくつかのモデルの下でベイズ推定が可能であることを確かめた。
- 適当な事前分布に対して新たな情報を加えることで事後分布が得られる。繰り返し情報を更新することで，より精度の高い事後分布が得られる。
- 事前分布や事後分布から，実際のデータの振舞いを予測することができる (予測分布)。
- モデルの適切さの指標として予測分布が利用できる。
- 母数同士が相関を持つ場合には同時分布の検証が必要。
- サンプリングのパラメータ (thinなど) は適宜調整すべし。

3章のもくじ

- ☑ 3.1 Inferring a rate
- ☑ 3.2 Difference between two rates
- ☑ 3.3 Inferring a common rate
- ☑ 3.4 Prior and posterior prediction
- ☑ 3.5 Posterior prediction
- ☑ 3.6 Joint distributions

参考文献

- 南風原朝和 (2002). 心理統計学の基礎――統合的理解のために　有斐閣アルマ
- 南風原朝和 (2014). 続・心理統計学の基礎――統合的理解を広げ深める　有斐閣アルマ
- 久保拓弥 (2012). データ解析のための統計モデリング入門――一般化線形モデル・階層ベイズモデル・MCMC　岩波書店
- Lee & Wagenmakers (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, U.K.: Cambridge University Press.
- 涌井良幸・涌井貞美 (2010). Excelでスッキリわかるベイズ統計入門　日本実業出版社