政治学方法論 I
第 3 回：統計モデル

矢内 勇生

大学院法学研究科・法学部

2016 年 4 月 27 日

神戸大学

今日の内容

1. 可能世界の分岐
   - 可能世界を数える
   - 事前情報を利用する
   - 数え上げから確率へ
2. 統計モデル
   - ベイズ統計分析
   - ベイズ統計モデルの構成要素
   - ベイズ統計モデルを使う

Garden of Forking Paths

Jorge Luis Borges. *Garden of Forking Paths.*

- 世界は可能性に満ちている
  - 法学部ではなく、経済学部を選んでいたら？
  - 政治学方法論 I を受講せず、その時間をアルバイトに使っていたら？
  - etc.
- 何かが起きる度に世界は分岐する
- 「今ここにある現実」とは異なる世界もあり得たはず

Garden of Forking Paths

Jorge Luis Borges. *Garden of Forking Paths.*

- 世界は可能性に満ちている
  - 法学部ではなく、経済学部を選んでいたら？
  - 政治学方法論 I を受講せず、その時間をアルバイトに使っていたら？
  - etc.
- 何かが起きる度に世界は分岐する
- 「今ここにある現実」とは異なる世界もあり得たはず
- 研究対象とする現象は、起こり得た世界の１つ
- その現象を生み出す経路が１つとは限らない

例題：袋の中のボールの色は？

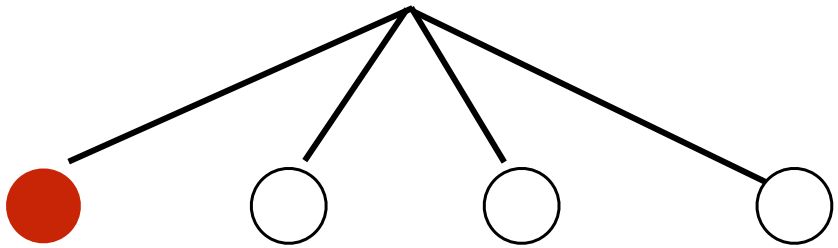
例題の設定

- 中身が見えない袋がある
- 袋の中に４つのボールが入っている
- 各ボールの色は、白または赤である
- 赤いボールと白いボールはそれぞれいくつ？
  - 仮説１ {赤 0, 白 4}
  - 仮説２ {赤 1, 白 3}
  - 仮説３ {赤 2, 白 2}
  - 仮説４ {赤 3, 白 1}
  - 仮説５ {赤 4, 白 0}
- 目標：どの仮説が最も妥当か判断する！

データ

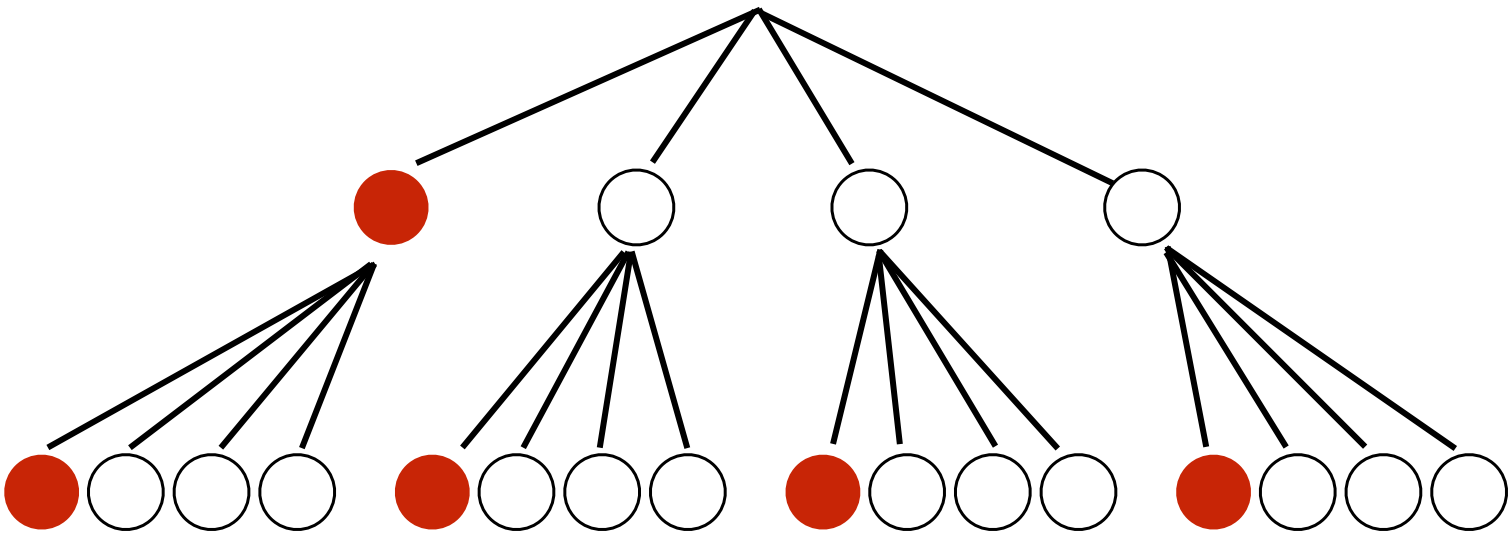
- データを手に入れる
  1. バッグを振って中身をよく混ぜる
  2. バッグからボールを 1 つ取り出して、ボールの色を記録する
  3. ボールをバッグに戻す
- 以上の過程を 3 回繰り返して得た結果：(赤, 白, 赤)
- データを利用して、どの仮説が最も妥当か考える

可能世界の検討：仮説 2 { 赤 1, 白 3}



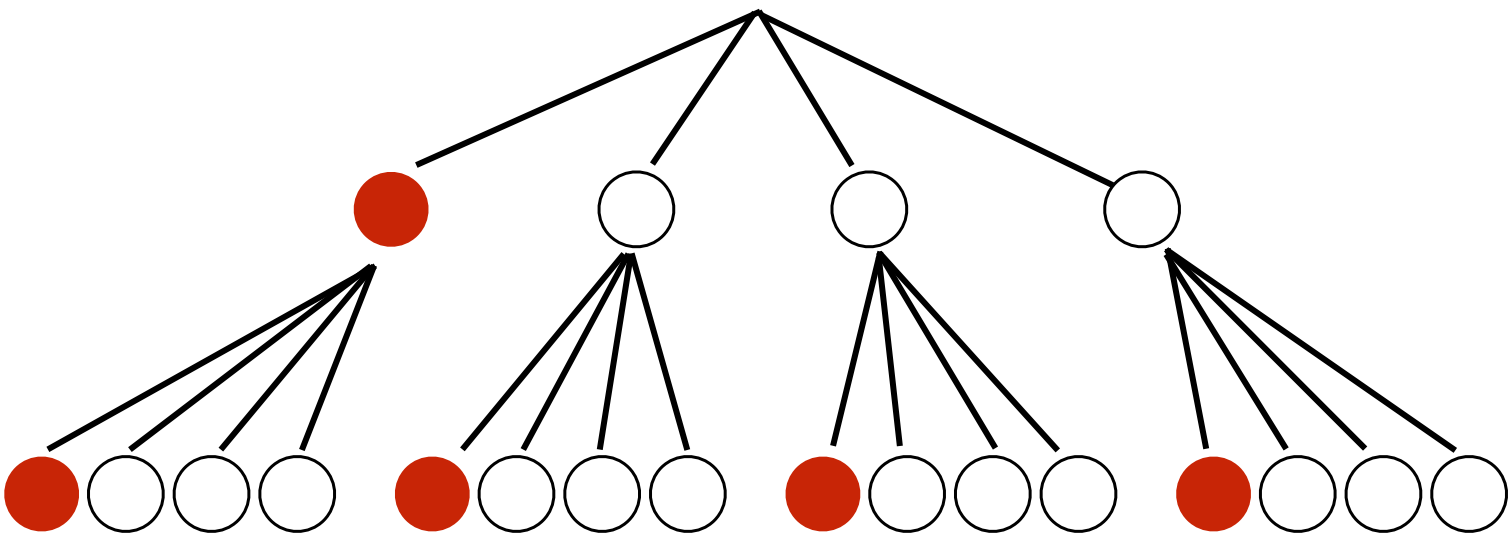
1 回目の分岐

可能世界の検討：仮説 2 { 赤 1, 白 3}



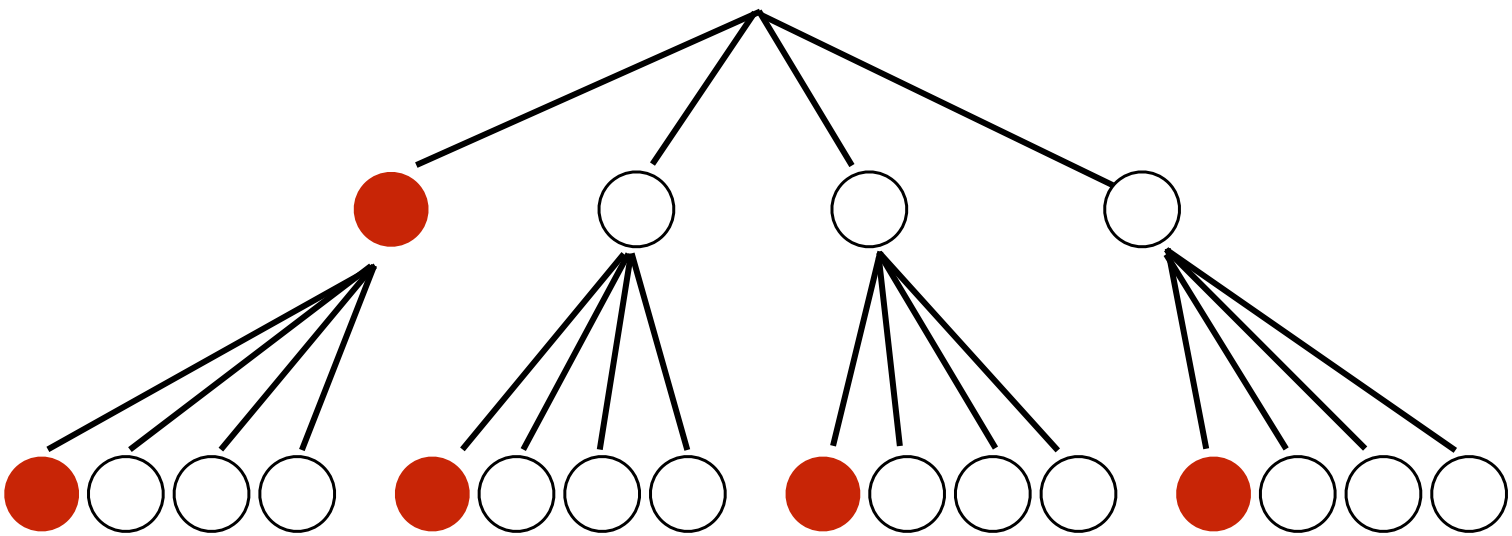
2 回目の分岐

可能世界の検討：仮説 2 { 赤 1, 白 3}



3 回目の分岐は？

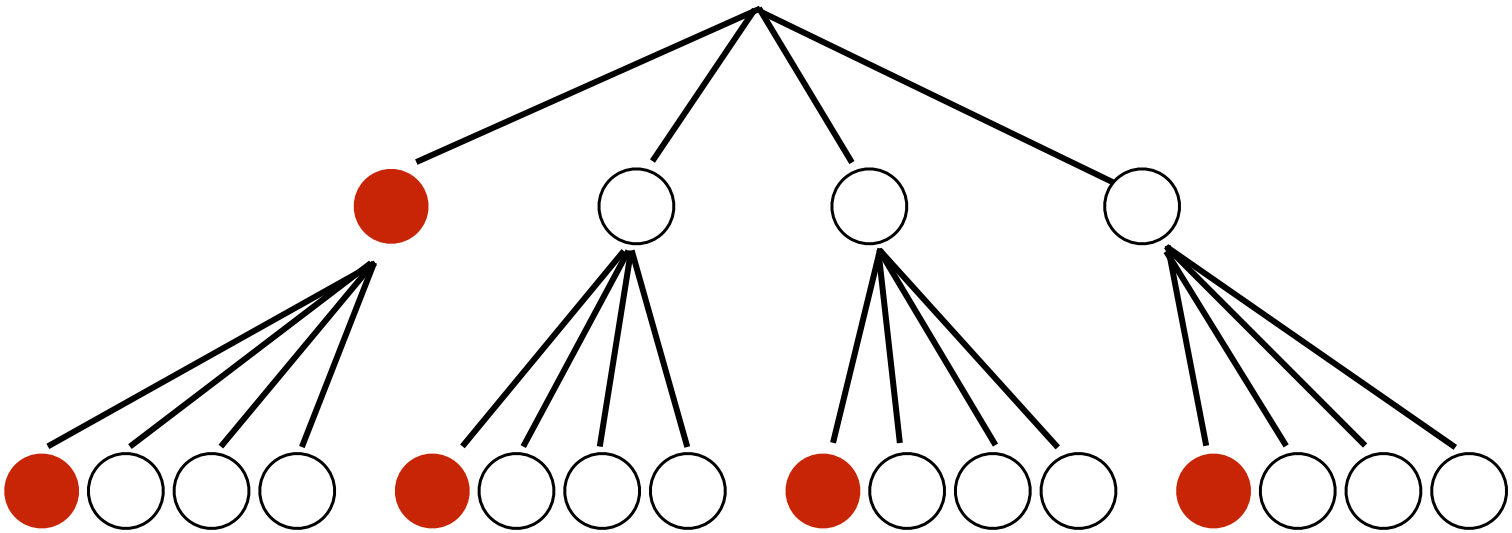
可能世界の検討：仮説 2 { 赤 1, 白 3}



3 回目の分岐は？

データ (赤, 白, 赤) に一致する経路はいくつある？

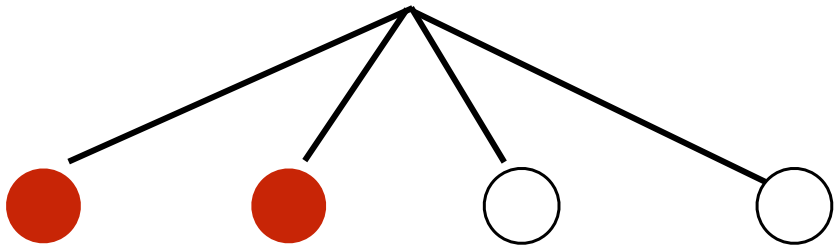
可能世界の検討：仮説 2 { 赤 1, 白 3}



3 回目の分岐は？
データ (赤, 白, 赤) に一致する経路はいくつある？

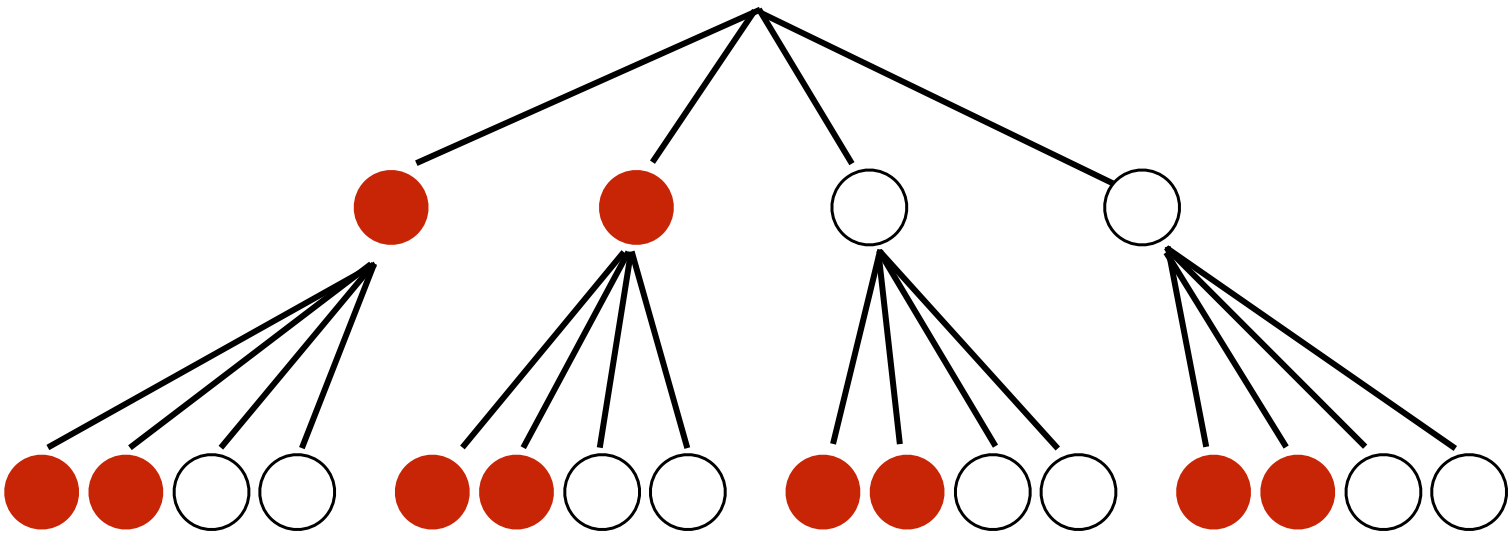
3 つ！

可能世界の検討：仮説 3 { 赤 2, 白 2}



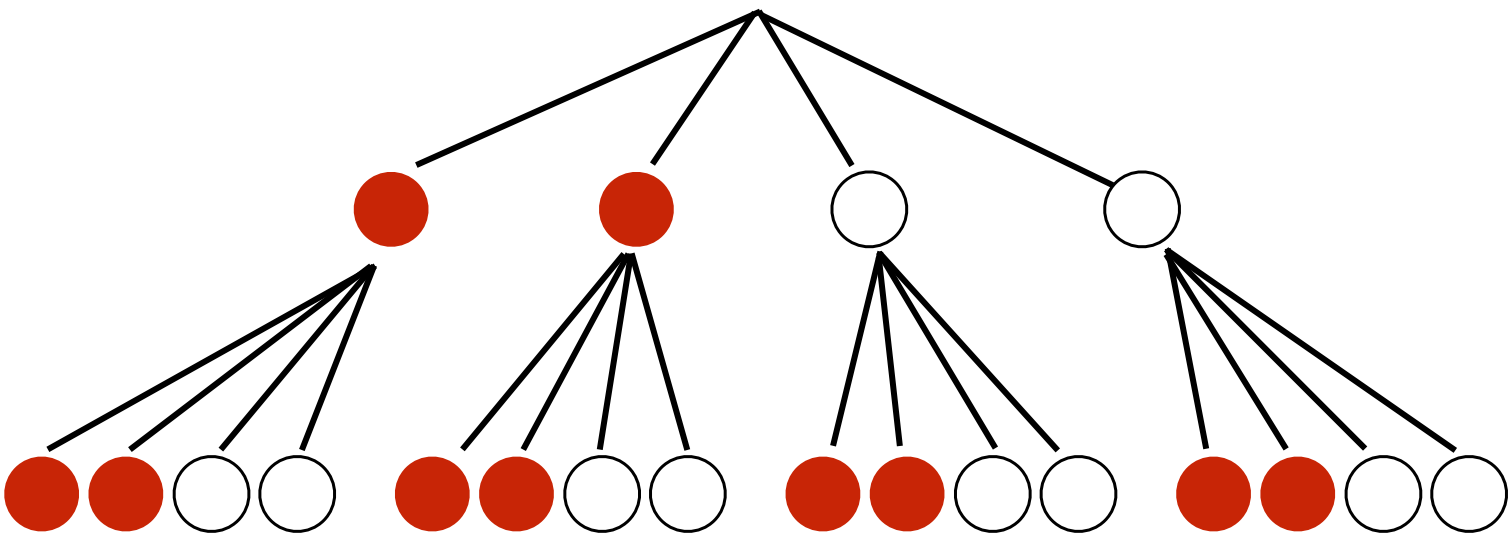
1 回目の分岐

可能世界の検討：仮説 3 { 赤 2, 白 2}



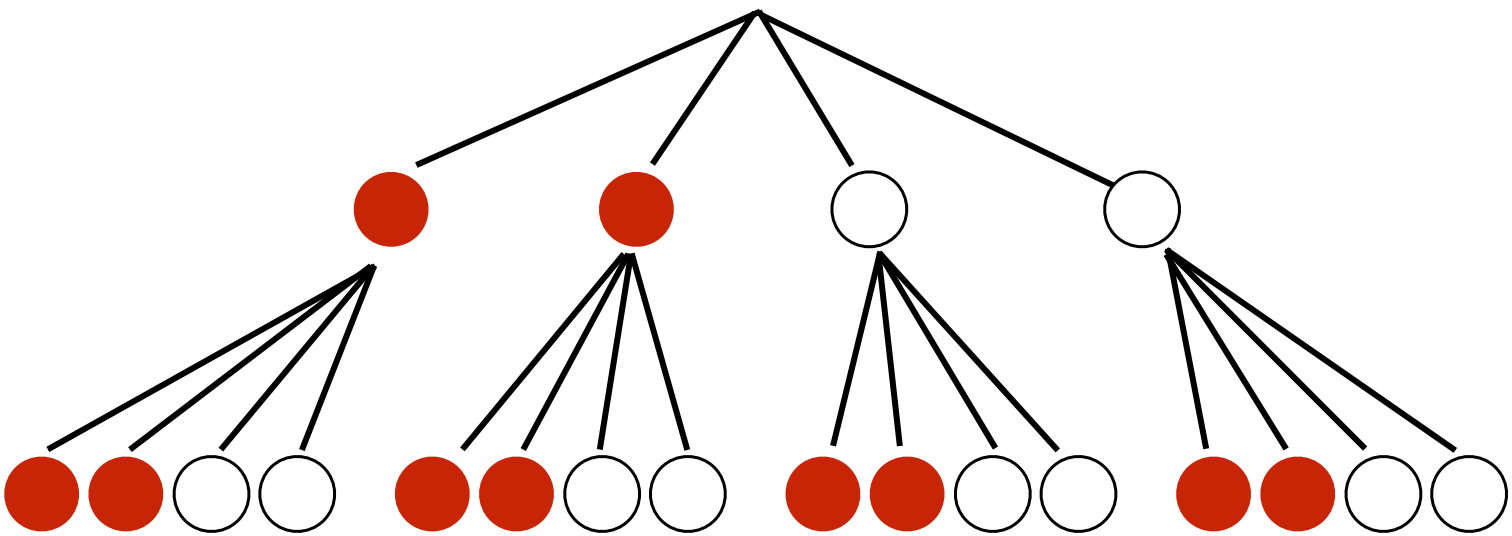
2 回目の分岐

可能世界の検討：仮説 3 { 赤 2, 白 2}



3 回目の分岐は？

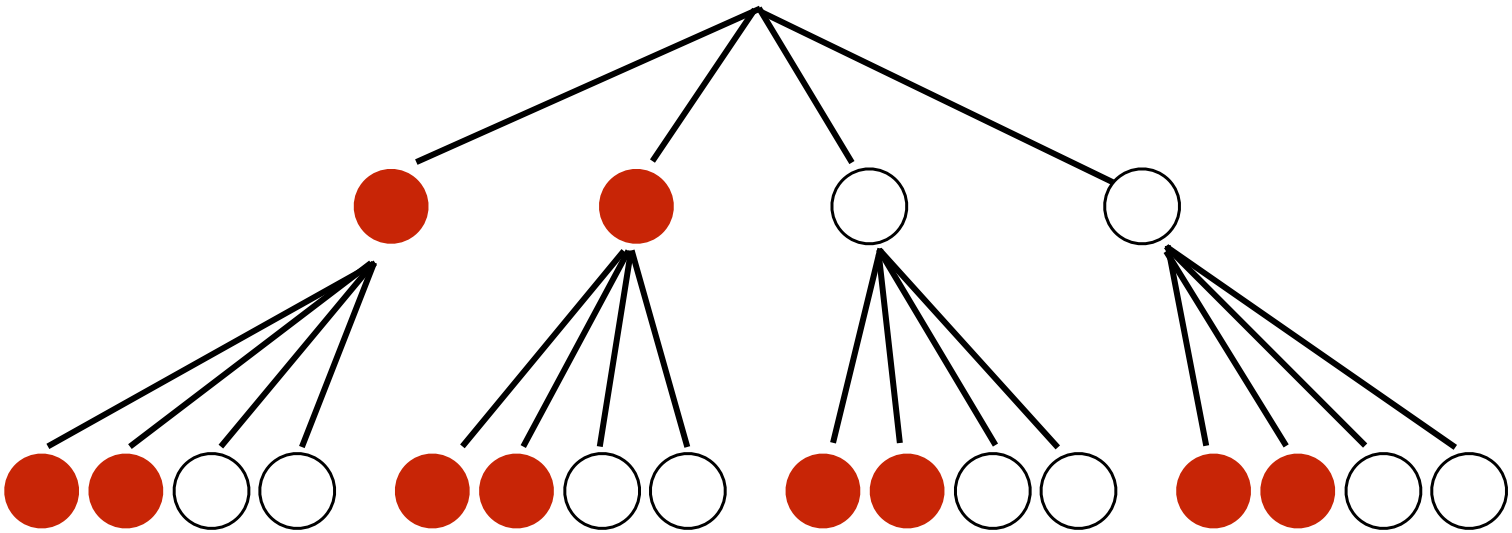
可能世界の検討：仮説 3 { 赤 2, 白 2}



3 回目の分岐は？

データ (赤, 白, 赤) に一致する経路はいくつある？

可能世界の検討：仮説 3 {赤 2, 白 2}



3 回目の分岐は？
データ (赤, 白, 赤) に一致する経路はいくつある？

8 つ！

可能世界の検討：観察されたデータを生み出す経路の数

仮説	仮説の内容	$\{$ 赤, 白, 赤 $\}$ を生み出す経路の数 (赤玉の数) $\times$ (白玉の数) $\times$ (赤玉の数)
1	(赤 0、白 4)	$0 \times 4 \times 0 = 0$
2	(赤 1、白 3)	$1 \times 3 \times 1 = 3$
3	(赤 2、白 2)	$2 \times 2 \times 2 = 8$
4	(赤 3、白 1)	$3 \times 1 \times 3 = 9$
5	(赤 4、白 0)	$4 \times 0 \times 4 = 0$

- 経路の数を比較して、どの仮説が最も妥当だと推論する？

可能世界の検討：観察されたデータを生み出す経路の数

仮説	仮説の内容	$\{$ 赤, 白, 赤 $\}$ を生み出す経路の数 (赤玉の数) $\times$ (白玉の数) $\times$ (赤玉の数)
1	(赤 0、白 4)	$0 \times 4 \times 0 = 0$
2	(赤 1、白 3)	$1 \times 3 \times 1 = 3$
3	(赤 2、白 2)	$2 \times 2 \times 2 = 8$
4	(赤 3、白 1)	$3 \times 1 \times 3 = 9$
5	(赤 4、白 0)	$4 \times 0 \times 4 = 0$

- 経路の数を比較して、どの仮説が最も妥当だと推論する？
- その推論にはどのような前提がある？
  - どの仮説の妥当性も、事前（データを見る前）には等しいという前提がある場合：仮説 4 が最も妥当
  - それ以外の場合は？

事前情報 (prior information) の利用

- 各仮説の相対的な妥当性について、事前に情報を持っていることがある
  - バッグに入れるボールの選び方を知っている
  - 過去の観察データを持っている：新しいデータを手に入れる度に、それまでのデータを事前の知識として使う

妥当性の更新

- 1 回目のボールの色を観察する前：すべての仮説の妥当性が同程度だとする
- (赤, 白, 赤) というデータが得られた
- この時点で、最も妥当なのは仮説 4 だが、仮説 3 の妥当性もそれに匹敵するくらい高い
- 新たなデータ：もう 1 度ボールを取り出したら、赤が出た
- この時点で、最も妥当な仮説はどれか？

データを事前情報と併せて使う

仮説	(赤) を生み出す経路の数	事前に数えた経路の数	新しい経路数
1.{ 赤 0, 白 4}	0	0	$0 \times 0 = 0$
2.{ 赤 1, 白 3}	1	3	$1 \times 3 = 3$
3.{ 赤 2, 白 2}	2	8	$2 \times 8 = 16$
4.{ 赤 3, 白 1}	3	9	$3 \times 9 = 27$
5.{ 赤 4, 白 0}	4	0	$4 \times 0 = 0$

- 経路数 = 事前の経路数 × 新しい経路数
- 掛け算：可能世界の分岐の数学的な表現
- 現時点で、仮説 4 の妥当性が最も高い
- 新しいデータと事前情報（古いデータ）を併せて利用したことで、仮説 4 の相対的な妥当性が高まった

異なる種類の情報を利用する（1）

- これまでの分析：同種の情報（同じ方法で行ったボールの色の観察）
- ボール入りバッグ製造・販売者からの情報提供：「赤玉は珍しい。ただし、ボール 1 色しかないバッグは販売しない」
  - 「赤玉 3 個入り 1 袋につき、赤玉 2 個入りは 2 袋、赤玉 1 個入りは 3 袋が流通するはずだ」
- これを新しいデータ、これまでの経路のカウントを事前情報として仮説の妥当性を再考する

異なる種類の情報を利用する（2）

仮説	新情報	事前情報	経路数
1.{ 赤 0, 白 4}	0	0	$0 \times 0 = 0$
2.{ 赤 1, 白 3}	3	2	$3 \times 2 = 6$
3.{ 赤 2, 白 2}	2	16	$2 \times 16 = 32$
4.{ 赤 3, 白 1}	1	27	$1 \times 27 = 27$
5.{ 赤 4, 白 0}	0	0	$0 \times 0 = 0$

- この時点で、仮説 3 の妥当性が最も高くなった
- 種類の異なる情報でも、推論に利用できる
- この分析から、どの仮説が最も妥当か結論を出せる？
- その結論は、どの程度「確か (certain)」か？

妥当性の更新法

- 「経路数」で妥当性を（ある程度）判断できる
- 経路数そのものは不便
  - それぞれの数字に意味はない：「32 vs 27」も「320 vs 270」も相対的妥当性は同じ
  - 「分岐」の回数が増えると、経路数がどんどん大きくなる

妥当性の更新法

- 「経路数」で妥当性を（ある程度）判断できる
- 経路数そのものは不便
  - それぞれの数字に意味はない：「32 vs 27」も「320 vs 270」も相対的妥当性は同じ
  - 「分岐」の回数が増えると、経路数がどんどん大きくなる
- 妥当性の更新法:
  - 仮説：H_i, $i \in \{1,2,3,4,5\}$
  - データ：$D =$(赤, 白, 赤)

$$D \text{を観察した後の} H_i \text{の妥当性}$$

$$\propto$$

$$H_i \text{が} D \text{を生み出す経路の数}$$

$$\times$$

$$H_i \text{の事前（} D \text{観察前）の妥当性}$$

仮説を数字で表す

- 袋に含まれる赤玉の割合を θ とする

  仮説 1 $\{$ 赤 0, 白 4$\} \rightarrow \theta = 0$
  仮説 2 $\{$ 赤 1, 白 3$\} \rightarrow \theta = 1/4 = 0.25$
  仮説 3 $\{$ 赤 2, 白 2$\} \rightarrow \theta = 2/4 = 0.5$
  仮説 4 $\{$ 赤 3, 白 1$\} \rightarrow \theta = 3/4 = 0.75$
  仮説 5 $\{$ 赤 4, 白 0$\} \rightarrow \theta = 4/4 = 1$

- $D =$(赤, 白, 赤)

D 観察後のθ_iの妥当性 $\propto$ θ_iの D への経路数 $\times$ θ_iの事前の妥当性

妥当性を標準化する

妥当性を標準化して確率にする

$$D\text{観察後の}\theta_i\text{の妥当性} = \frac{\theta_i\text{の}D\text{への経路数} \times \theta_i\text{の事前の妥当性}}{\sum_i(\theta_i\text{の}D\text{への経路数} \times \theta_i\text{の事前の妥当性})}$$

仮説	赤玉の割合 θ	経路数	妥当性（確率）
{赤 0, 白 4}	0.00	0	0 / 20 = 0
{赤 1, 白 3}	0.25	3	3 / 20 = 0.15
{赤 2, 白 2}	0.50	8	8 / 20 = 0.4
{赤 3, 白 1}	0.75	9	9 / 20 = 0.45
{赤 4, 白 0}	1.00	0	0 / 20 = 0
	合計	20	1

妥当性を標準化する

妥当性を標準化して確率にする

$$D\text{観察後の}\theta_i\text{の妥当性} = \frac{\theta_i\text{の}D\text{への経路数} \times \theta_i\text{の事前の妥当性}}{\sum_i(\theta_i\text{の}D\text{への経路数} \times \theta_i\text{の事前の妥当性})}$$

仮説	赤玉の割合 θ	経路数	妥当性（確率）
{赤 0, 白 4}	0.00	0	0 / 20 = 0
{赤 1, 白 3}	0.25	3	3 / 20 = 0.15
{赤 2, 白 2}	0.50	8	8 / 20 = 0.4
{赤 3, 白 1}	0.75	9	9 / 20 = 0.45
{赤 4, 白 0}	1.00	0	0 / 20 = 0
	合計	20	1

確率を使うことによって、推論しやすくなる！

専門用語の導入

- θ のように、推定の対象となる数（仮説の中身を構成するもの）：母数（パラメタ、**parameter**）
- ある仮説が特定のデータを生み出す経路の相対的な数の大きさ：尤度 **(likelihood)**
- データを観察する前の時点での特定の θ の妥当性：事前確率 **(prior probability)**
- データを使って更新された特定の θ の妥当性：事後確率 **(posterior probability)**

例題

地球儀問題：地球表面の水の割合は？ (McElreath 2016: Ch.2)

手の平サイズの地球儀がある。地表のうち、水に覆われている割合を知りたい。そこで、次の方法で調べることにした。地球儀を投げ上げ、落ちてきた地球儀を両手でキャッチする。そのとき、右手人差し指の先端部が触れているのが水 (W) か陸地 (L) を記録する。この作業を何度か繰り返す。作業を 9 回実行した結果、

$$D = (W, L, W, W, W, L, W, L, W,)$$

というデータが得られた。

ベイズ統計モデルによる推論

1. データがどのように生み出されたかを考える
2. データを使ってモデルを「教育」する
3. モデルを評価する

データ生成過程 (Data Generating Process)

- データ生成過程 (data generating process: DGP) を考える
  - 記述的なモデル (descriptive model)
  - 因果モデル (causal model)
- 問題の背景にある事実と、データがどのように収集、観察されたかを考慮に入れる
- 地球儀問題の場合
  - 地球儀表面の水の本当の割合：θ
  - 地球儀を 1 回投げ上げたとき、W が観察される確率は θ、L が観察される確率は $1-\theta$
  - 地球儀の投げ上げを繰り返すとき、各投げ上げは互いに独立
- データ分析のために、DGP を数式で表現する

ベイズ更新 (Bayesian Updating)

複数の仮説のうち、どの仮説が相対的に妥当かを判断したい

- データ分析前
  - （事前の）妥当性をもっている
  - データ生成過程をモデル化する
- データによって、情報を更新する（学習）
- 更新の仕方はモデルに依存する
- 追加のデータを手に入れたら、さらに情報を更新する

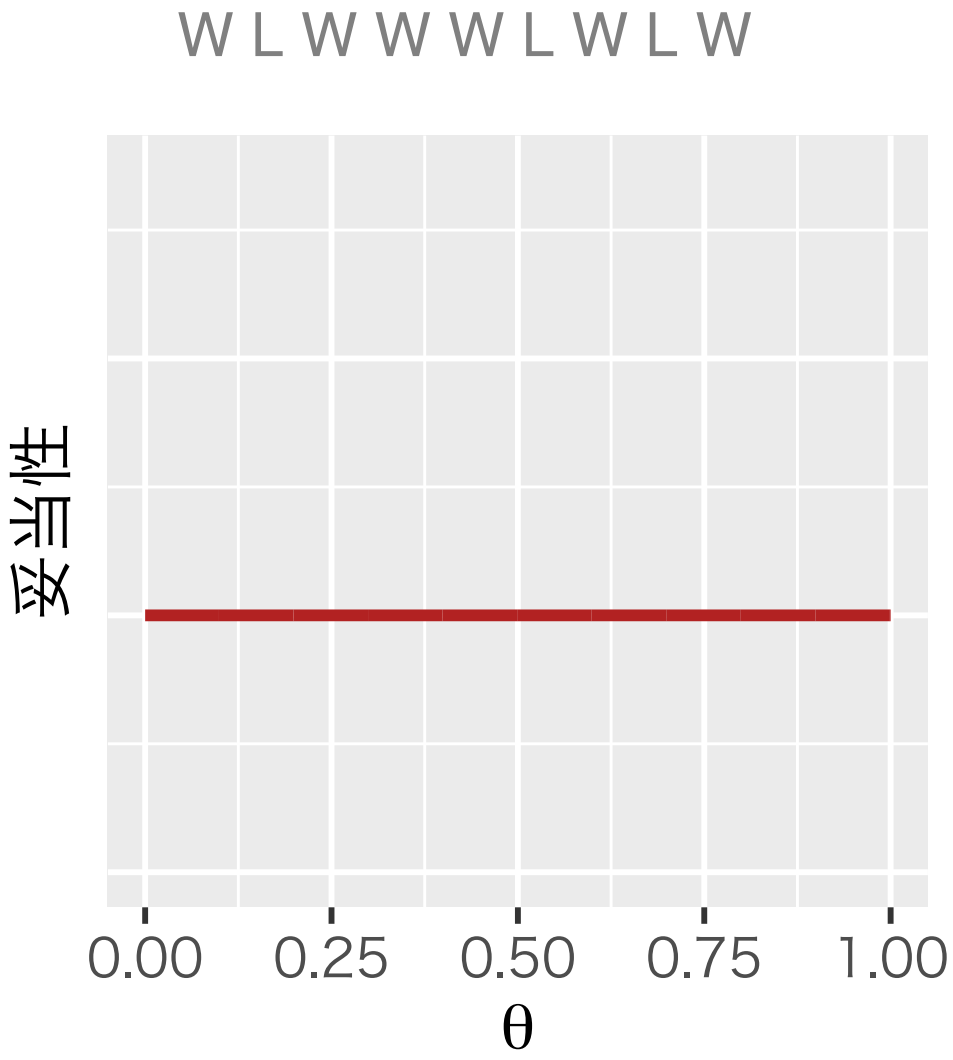


図: $n = 0$

W L W W W L W L W

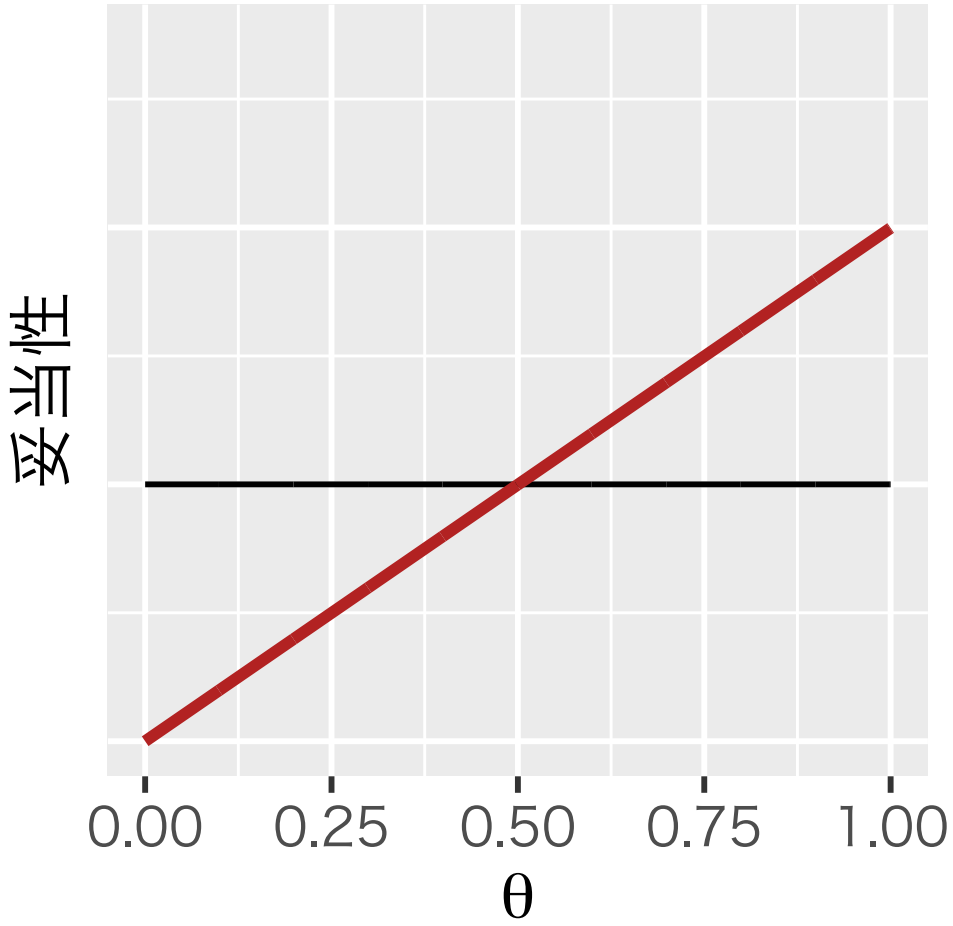


図: $n = 1$

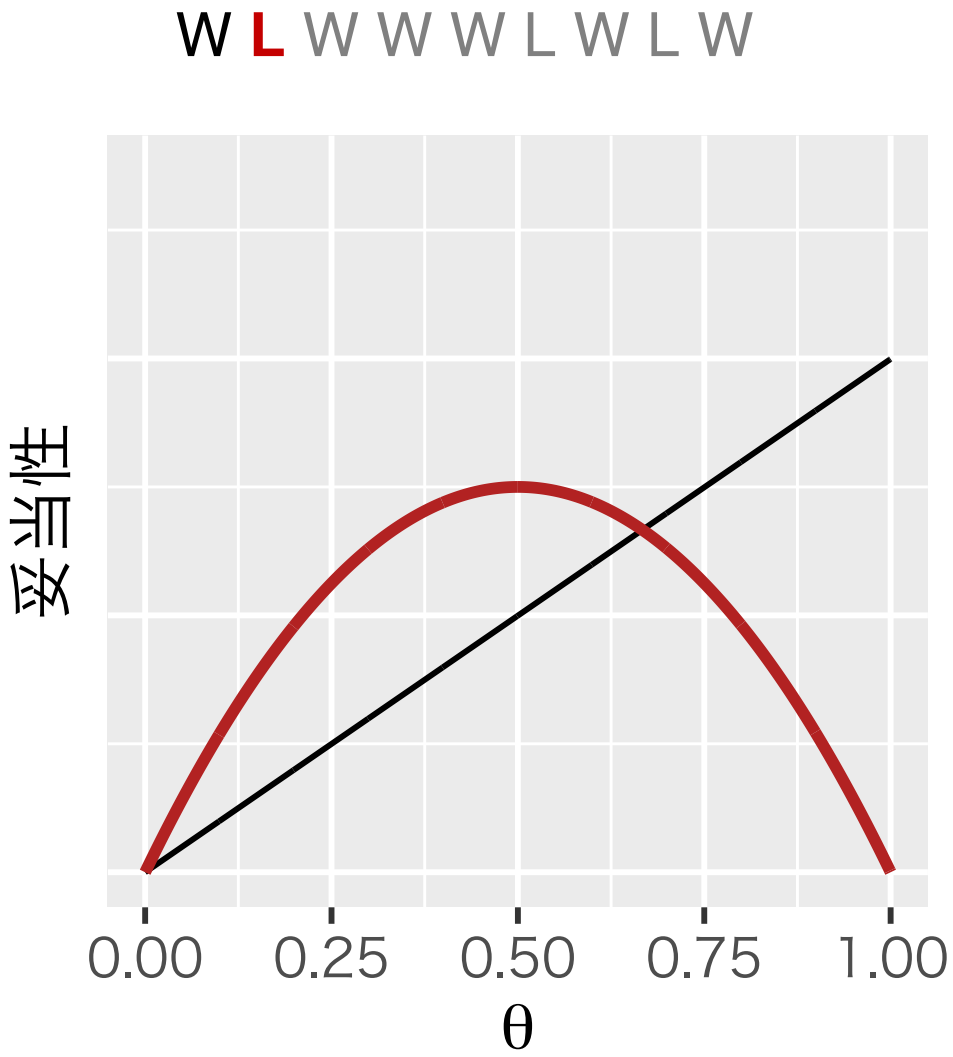


図: $n = 2$

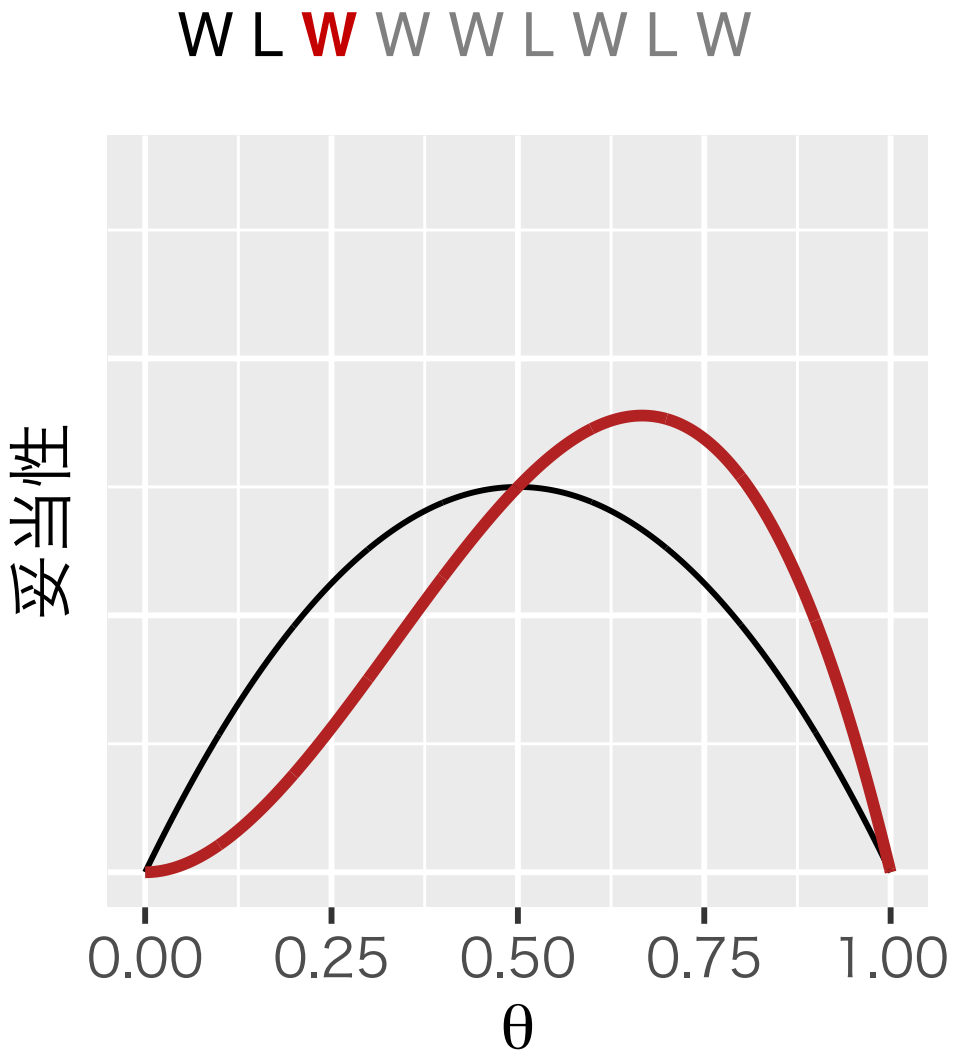


図: $n = 3$

W L W **W** W L W L W

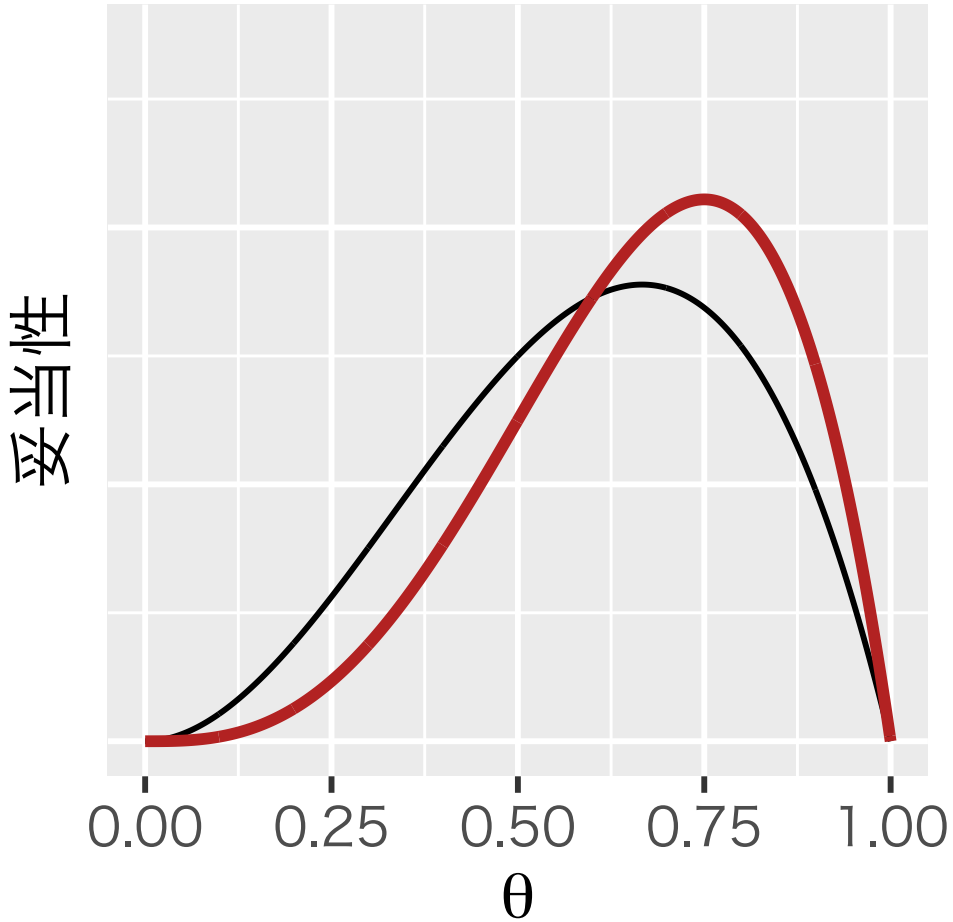


図: $n = 4$

W L W W **W** L W L W

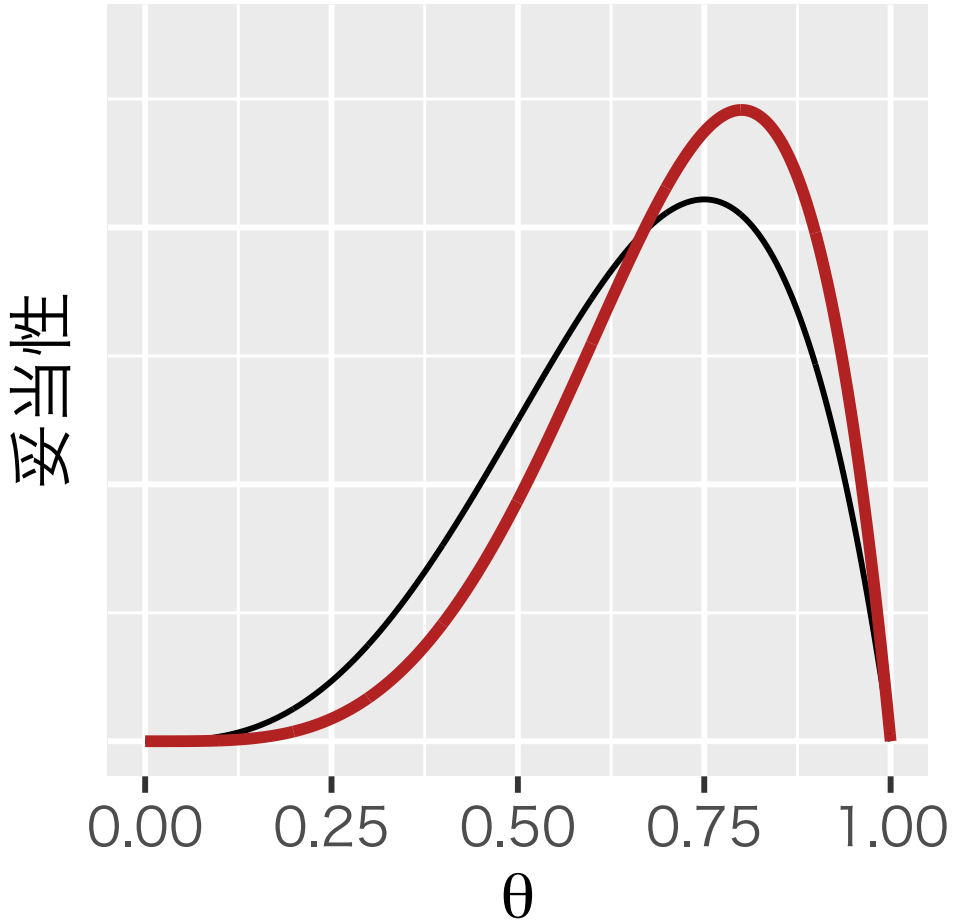


図: $n = 5$

W L W W W L W L W

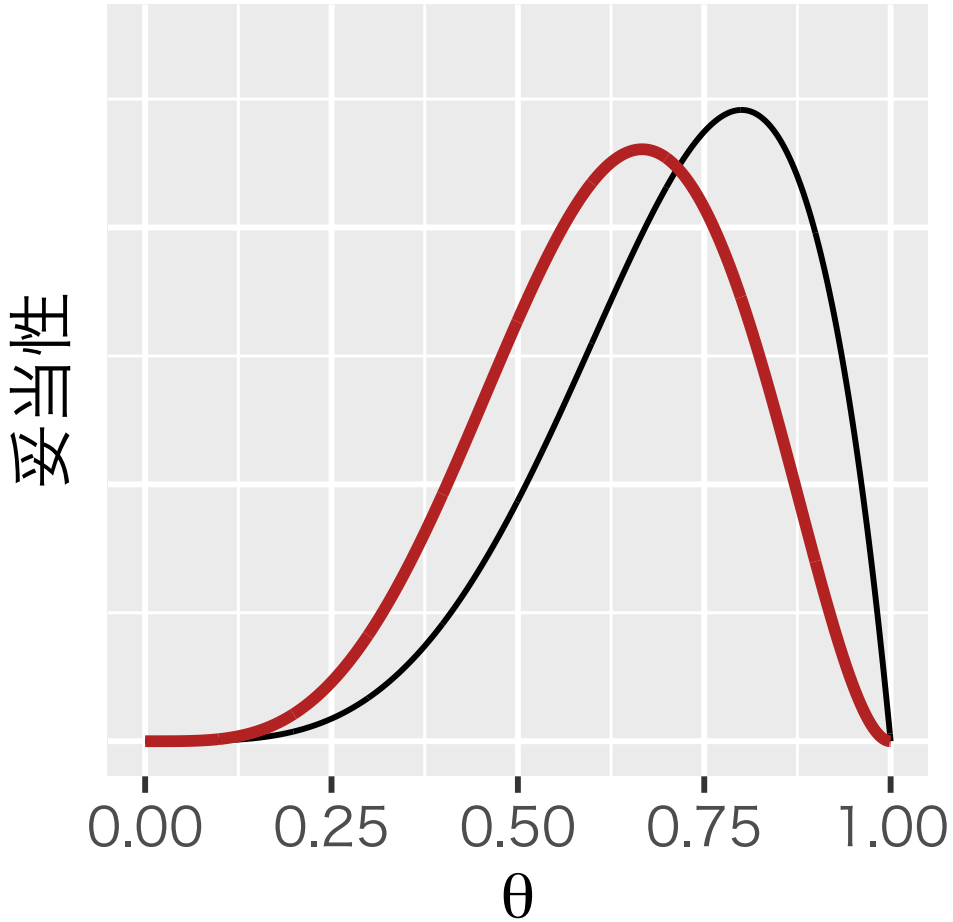


図: $n = 6$

W L W W W L W L W

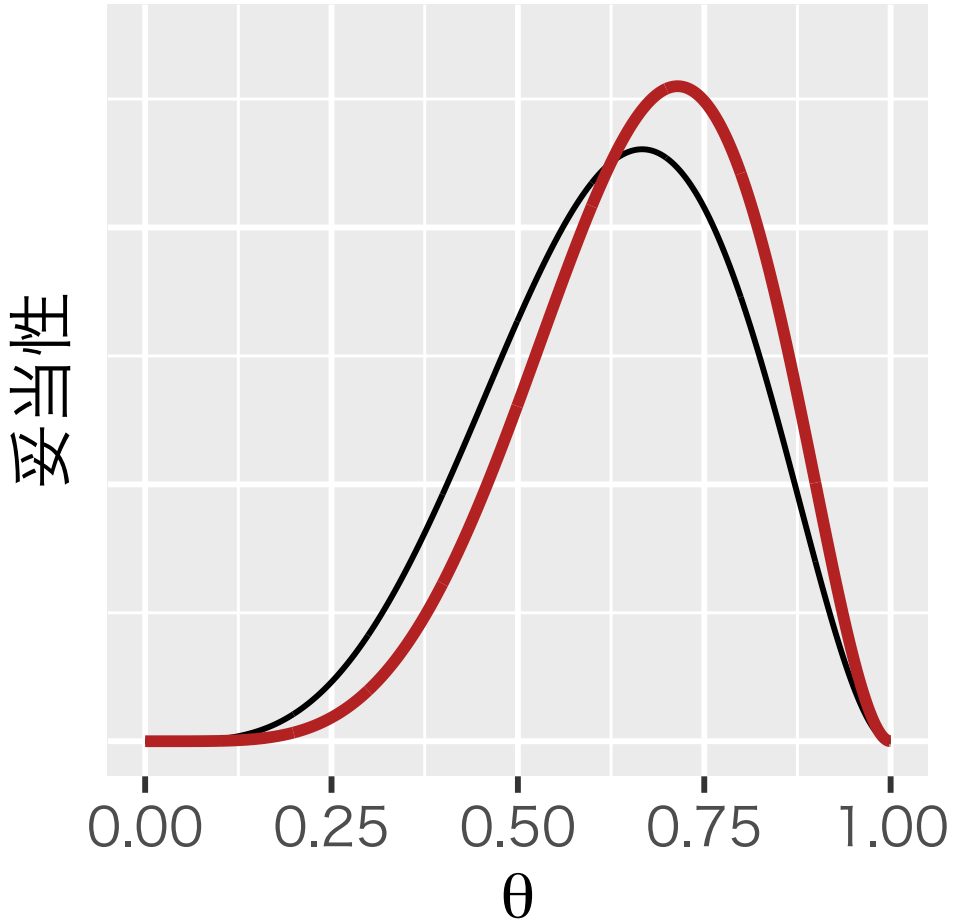


図: $n = 7$

W L W W W L W L W

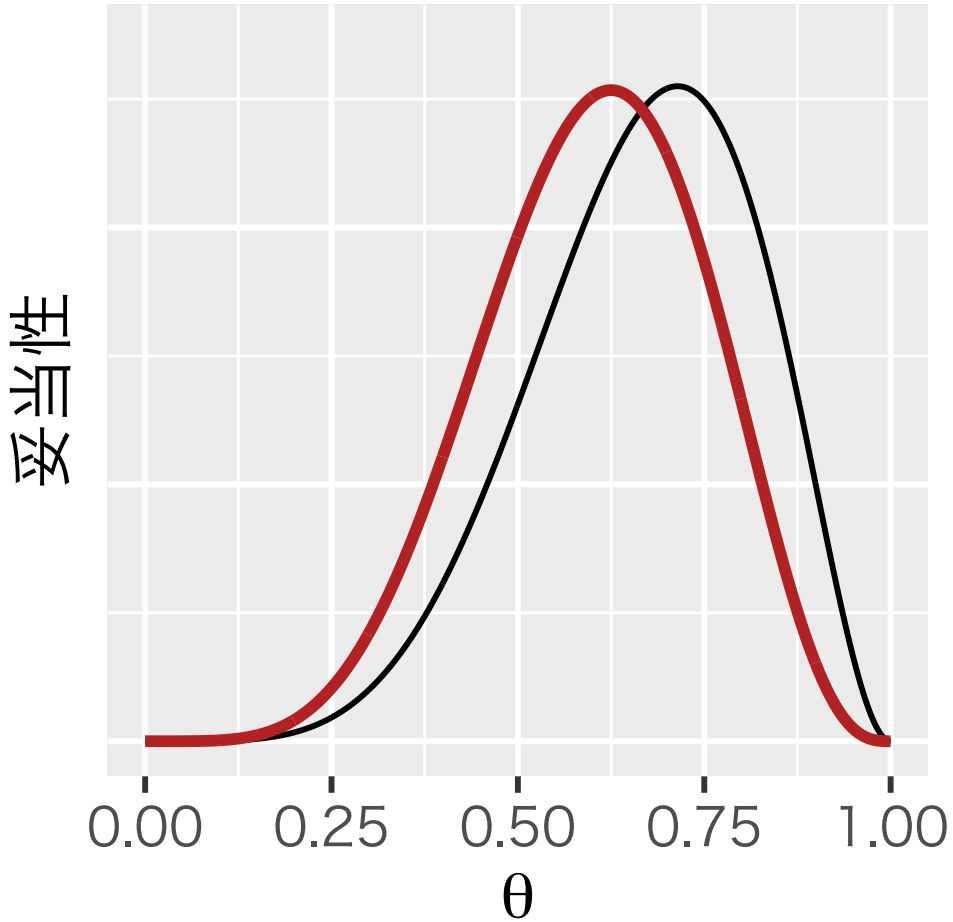


図: $n = 8$

W L W W W L W L **W**

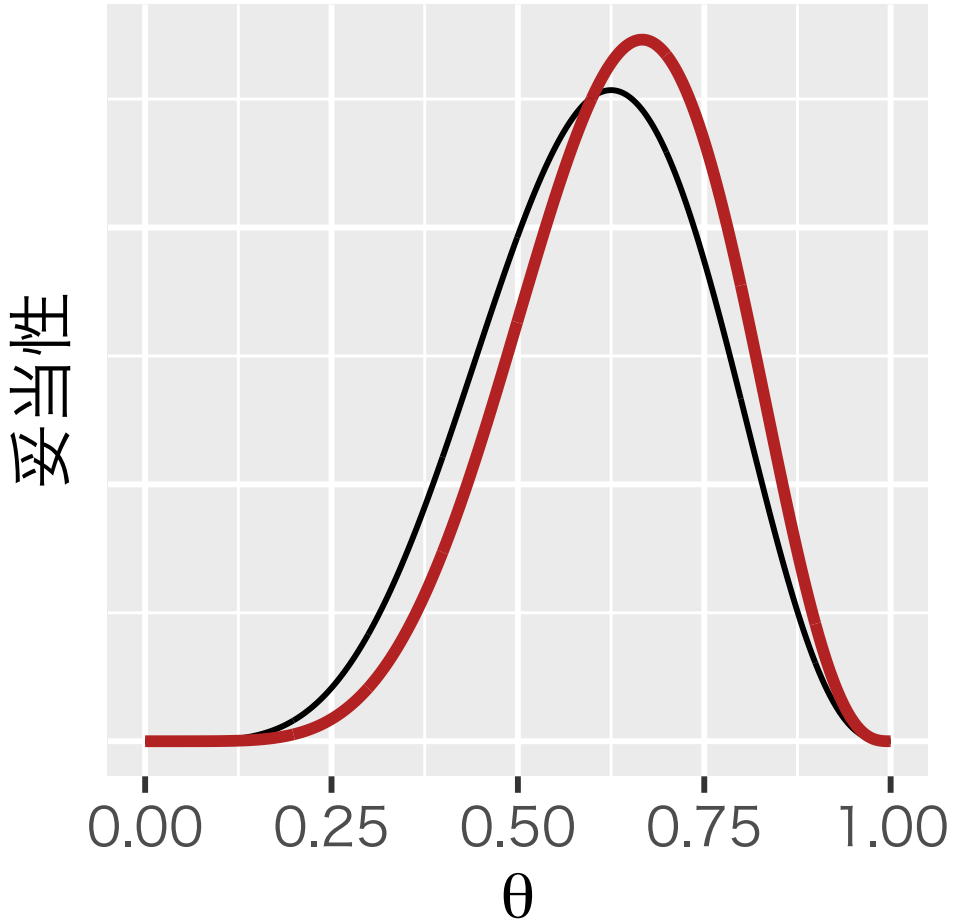


図: $n = 9$

モデルを評価する

ベイズ更新が終わっても、結果を過信せず、批判的に評価することが必要

- 推定の不確実性が低いモデルが正しいとは限らない
  - 観測数 n が大きくなれば、誤ったモデルでも「精度の高い」（誤った）推定結果を示す
  - 推定値はモデルに依存する：モデルを変えれば、推定値は変わり得る
- モデルが重大な見落としをしていないか考える
  - 厳密には、モデルの仮定はほぼ常に「正しく」ない
  - モデルの誤りが見つけられない $\neq$ モデルが正しい
  - 結果を変え得るような要因がモデルから抜け落ちていないか？
  - 一見すると瑣末な仮定の変更が、結果を大きく変える可能性はないか？
  - 例：データ内の W と L の順序は結果を左右するか？

ベイズ統計モデルの構成

データを分析する前に、以下の３つを用意する

1. 尤度関数
2. １つ以上の母数
3. 事前確率

- 通常は、この順番で用意する
- これらの要素とデータを利用して、相対的に妥当性が高い仮説を見つける

尤度 (Likelihood)

- 特定のデータがある仮説から生み出される妥当性を表す数式
- 各仮説について、相対的な妥当性を数字で表すことを可能にする
- データ生成過程に適した尤度関数を作る
  - 自分で尤度関数を定義する
  - よく使われる尤度関数を利用する

地球儀問題の尤度

- データ生成過程
  - 地球儀を投げ上げた結果：W または L
  - それぞれの投げ上げは独立
  - 水の割合 θ はどの投げ上げでも同じ
  - データ：投げ上げを n 回行うと、w 回 W が出る

地球儀問題の尤度

- データ生成過程
  - 地球儀を投げ上げた結果：W または L
  - それぞれの投げ上げは独立
  - 水の割合 θ はどの投げ上げでも同じ
  - データ：投げ上げを n 回行うと、w 回 W が出る
- 二項分布 (binomial distribution)

$$w|n,\theta \sim \mathrm{Bin}(n,\theta)$$

$$\Pr(w|n,\theta) = \binom{n}{w}\theta^w(1-\theta)^{n-w} = L(\theta|w,n)$$

母数（パラメタ, Parameter）

- 尤度関数の中で、異なる値を取り得るもの：二項分布の場合 n, w, θ
- これらのうち、どれか１つ以上を母数として扱う：問題によって異なる
- 地球儀問題の場合
  - n と w はデータとして観察される
  - 直接観察されない θ を母数とする
- 母数：データ分析における推定の対象
- ギリシャ文字 $(\alpha, \beta, \gamma, \ldots)$ で表されることが多い

事前確率 (Prior)

- ベイズ統計分析：すべての母数に事前確率を与える
- 事前確率：データ観察前の時点で、母数が取り得る値のそれぞれが、相対的にどの程度妥当かを表す
- 地球儀問題の場合
  - 母数 θ は割合：$\theta \in [0, 1]$
  - 区間 $[0, 1]$ の各値が、どの程度妥当と言えるか表す必要がある
  - $[0, 1]$ の範囲で妥当性に差がない（妥当性が等しい）場合：一様分布で事前確率を表す

$$\theta \sim \text{Unif}(0, 1)$$

$$\Pr(\theta) = \frac{1}{1-0} = 1$$

- 事前確率の選び方は様々

事後確率 (Posterior)

- 尤度と事前確率が決まったら、ベイズの定理を用いて事後確率を求める

ベイズの定理 (Bayes Theorem)

$$\begin{aligned}\Pr(\theta|w) &= \frac{\Pr(w|\theta)\Pr(\theta)}{\Pr(w)} \\ &= \frac{L(\theta|w)\Pr(\theta)}{\int \Pr(w|\theta)\Pr(\theta)d\theta}\end{aligned}$$

$$\text{事後確率} = \frac{\text{尤度} \times \text{事前確率}}{\text{尤度の平均値}} \propto \text{尤度} \times \text{事前確率}$$

地球儀問題の事後確率の一例*

- 尤度：$\Pr(w|n,\theta) = \frac{n!}{w!(n-w)!}\theta^w(1-\theta)^{n-w}$
- 事前確率：$\theta = 1$

$$\Pr(\theta|w,n) \propto \Pr(w|n,\theta) \propto \theta^w(1-\theta)^{n-w} \times 1$$
$$= \Pr(w|n,\theta) \propto \theta^w(1-\theta)^{n-w}$$

$$\theta|n,w \sim \mathrm{Beta}(w+1, n-w+1)$$

事後確率の導出

- 問題が簡単なとき：解析的に答えを出せる
  - 地球儀問題はこれが可能
- 問題が複雑になると、解析的に答えを出すのが大変か実質的に不可能
  - 母数の数が多いとき
  - 尤度関数が複雑なとき
- 近似的に答えを出す
  1. グリッド近似
  2. 二次近似
  3. マルコフ連鎖モンテカルロ法

グリッド近似 (Grid Approximation)

母数のグリッド（格子）を作り、各グリッドで事後確率を計算する

長所　事後確率を導出する仕組みがよくわかる（教育的）

短所　母数が増えると使えない（非実践的）

グリッド近似のプロセス（詳細は web 資料を参照）

1. グリッドを定義する
2. グリッド上の各点に、事前確率を与える（計算する）
3. グリッド上の各点で、尤度を計算する
4. 標準化されていない事後確率を計算する
5. 事後確率を標準化する

二次近似 (Quadratic Approximation)

- やや難しいので、とりあえず使わない
- 必要になったときに説明する
- 一言で言えば、事後分布を放物線（二次関数）に単純化して考える方法
- こういう方法があるということは知っておいて欲しい

マルコフ連鎖モンテカルロ法 (MCMC)

- Markov chain Monte Carlo (MCMC)
- 実践的には、よく使われる方法
- 政治学方法論 II で詳しく説明する予定
- こういう方法があるということは知っておいて欲しい