Bayesian Data Analysis : Chapter2 Single-parameter models

Andrew Gelman, John B.Carlin, Hal S.Stern,
David B.Dunson, Aki Vehtari, and Donald B.Rubin

山崎 遼也

情報学科 数理工学コース 4 回

2017/3/28～4/1

Table Contents

Key Words

- ベルヌーイ試行
- ラプラスの連続法則
- 不十分な理由の原理
- 超パラメータ
- 共役
- 指数型分布族
- 十分統計量
- 参照事前分布
- 適切, 不適切
- Jeffry の不変原理
- ピボット量
- 無情報, 弱情報事前分布

ベイズ推論の最初の詳細な議論は, 単一のスカラパラメータのみが推定される統計モデルの文脈である. すなわち, 推定値 θ は 1 次元である. この章では, 二項, 正規, ポアソン, 指数の 4 つの基本的かつ広く使用される 1 次元モデルを考察すると同時に, ベイズデータ分析の重要な概念と計算方法を紹介します.

2.1 二項データからの確率の推定

単純な二項モデルでは, 目標はベルヌーイ試行 (Bernoulli trails) の系列の結果, すなわち, それぞれが 0 か 1 であるようなデータ $y_1, \cdots, y_n$ から未知の母集団の割合を推定することである. この問題は, 比較的単純であるが重要なベイズ推論の議論の出発点である. 二項モデルで始めることにより, 私たちの議論は, 1763 年に Thomas Bayes によって最初に出版されたベイズ分析と並行しており, 彼の精神的貢献は依然として興味深い.

二項分布はそれぞれの試行が 2 つの, 伝統的には成功 (success) と失敗 (failure) とラベル付される結果を取りうる n 個の交換可能な試行または大きな母集団化からの抽出であるデータに対して自然なモデルを提供する. 交換可能性より, データは n 回の試行における成功の総数で要約できる. これを, ここでは y と表す. 独立して同一分布の確率変数を用いた交換可能な試行についての公式からの公式への変換は, パラメータ θ を母集団における成功の割合, または同等に, 各試行における成功の確率を表すことによって自然に達成される. 二項サンプリングモデルは,

$$p(y|\theta) = \mathrm{Bin}(y|n,\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \tag{2.1}$$

である. 一定である実験計画の一部と見るため, 左辺で n 依存性を抑える. この問題のすべての確率は, n に対して条件付きであると仮定する.

例. 女児が出生する確率の推定

二項モデルの具体的な応用として, 出生人口のうちで性比の推定を考える. 女性である出生の割合は, 科学的にも一般にも興味深いトピックである. 200 年前, ヨーロッパ人口における女性出生の割合が 0.5 未満であることが確立された (以下の史料を参照) が, 今世紀には性比に影響を与える要因に焦点を当てている. 大きなヨーロッパ人口での女性出生の最近認められた値は 0.485 である. この例に対して女性出生の割合に対してパラメータ θ を定義するが, このパラメータを報告する代わりの方法は, 女性に対する男性の出生比, $\phi = (1-\theta)/\theta$ である. n 人出生した中で, y を女の子の数とする. 二項モデル (2.1) に訴えることで, n 人の出生は θ を条件として独立していると仮定しており, すべての場合において θ と等しい確率で出生する. このモデル仮定は赤ちゃんの性別に影響を与えるであろう (例えば, 同じ家族内の複数の出生または出生と区別する) 追加の情報がないときそうであると判断される交換可能性に選るものである.

二項モデルでベイズ推論を行うために, θ に対する事前分布を特定しなければならない. この本を通して何度も事前分布の特定に関する問題について議論するだろうが, この時点では簡単のため, θ に対する事前分布を区間 $[0, 1]$ での一様分布と仮定する.

(2.1) に適用された (1.2) で示されたベイズ則の適用は, θ の事後密度を

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} \tag{2.2}$$

として与える. 定数 n, y に関して, $\binom{n}{y}$ は未知パラメータ θ に依存しない. 多くの例と典型的であるように, 事後確率はすぐに閉じた形で書ける. 1 パラメータ問題では, これより事後分布の図的な説明ができる. 例えば, 図 2.1 では, いくつかの異なる実験, すなわち n, y の異なる値について, 正規化されていない密度 (2.2) が表示されている. 4 つの実験のそれぞれは成功率が同じですが, サンプルサイズは異なります. 現在の場合では, (2.2) をベータ分布の非正規された形として認識できる.(付録 A を参照)

$$\theta|y \sim \text{Beta}(y+1, n-y+1) \tag{2.3}$$

歴史的メモ: Bayes と Laplace

多くの初期の確率論の著者は, 基本の二項モデルを扱っていた. 17 世紀と 18 世紀初頭の永続的な意義の最初の貢献は, 事前データという問題に集中していました. θ を与えられたとき, 確率変数 y の様々な可能な結果の確率はどれくらいですか? 例えば, Jacob Bernoulli の対数の弱法則は, 任意の θ と任意の $\epsilon > 0$ について, $y \to \mathrm{Bin}(n, \theta)$ ならば, $n \to \infty$ とすると $\Pr(|\frac{n}{y} - \theta| > \theta) \to 0$ であるということを主張する. 彼の生涯に出版されなかった英国の数学者トーマス・ベイズ (Thomas Bayes) と, フランスのナポレオン時代に広がった創造的で生産的な数学者のピエール・サイモン・ラプラスは, 確率を逆転させる最初の独立クレジットを受け取る観測された y について, θ についての確率表記を得る. 1763 年に出版された, 彼の有名な論文で, ベイズは $\Pr(\theta \in (\theta_1, \theta_2)|y)$ を考えた. 彼の解は, 確率空間の物理的な類推に基づいて, (ビリヤード台のような) 長方形のテーブルを作成しました.

1. (事前分布) ボール W は無作為に投げられる (テーブル上の一様な分布に従って). テーブル上のボールの水平位置は θ であり, テーブル幅の分数として表されます.

2. (尤度) ボール O は無作為に n 回投げられる. y の値は, O が W の右に掛かる回数です.

すると, θ は [0, 1] 上の (事前) 一様分布を持つと仮定される. 論文で導かれた直接の確率計算を用いて, ベイズは

$$\begin{aligned}\Pr(\theta \in (\theta_1, \theta_2)|y) &= \frac{\Pr(\theta \in (\theta_1, \theta_2), y)}{p(y)} \\ &= \frac{\int_{\theta_1}^{\theta_2} p(y|\theta)p(\theta)d\theta}{p(y)} \\ &= \frac{\int_{\theta_1}^{\theta_2} \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}}{p(y)} \end{aligned} \tag{2.4}$$

ベイズは分母を計算することにも成功した.

$$\begin{aligned} p(y) &= \int_0^1 \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} d\theta \\ &= \frac{1}{n+1} \text{ for } y = 0, \cdots, n \end{aligned} \tag{2.5}$$

この計算は, y のすべての可能な値が同様に先験的に可能性が高いことを示す. (2.4) の分子は, $y, (n-y)$ の大きな値に対して閉形式の式を持たない不完全なベータ積分であり, ベイズにとっては明らかにいくつかの困難を示した.

しかし, ラプラスは独自にベイズ定理を発見し, 積分計算のための新しい解析ツールを開発しました. 例えば, 彼は関数 $\theta^y(1-\theta)^{n-y}$ を $\theta = y/n$ でその最大値の周りに拡張し, 現在の正規近似として未知ベータ積分を評価した.

二項モデルの解析では, ラプラスも一様事前分布を用いた. 彼の最初の深刻な適用は, 人口における少女の出生の割合を推定することでした. 合計で 241,945 人の少女と 251,527 人の少年がパリで 1745 年から 1770 年に生まれました. θ をある出生が女性である確率とすると, ラプラスは

$$\Pr(\theta \geq 0.5 | y = 241,945, n = 251.527 + 241,945) \simeq 1.15 \times 10^{-42}$$

と $\theta < 0.5$ が確かであるということを示した.

予測

一様事前分布を持つ二項の例では, 既に (2.5) で述べたように, 事前予測分布を明示的に評価することができます. モデルの下では, y のすべての可能な値は, 同様に, 先験的に可能性が高い. このモデルからの事後予測に対して, 新しい n 個の試行の他の集合よりも, 新しい 1 つの試行の結果により興味があるかもしれない. $\tilde{y}$ をはじめの n 個と交換可能な, 新しい試行の結果とすると, ベータ分布の特性 (付録 A を参照) から

$$\begin{aligned}\Pr(\tilde{y}=1|y) &= \int_0^1 \Pr(\tilde{y}=1|\theta,y)p(\theta|y)d\theta \\ &= \int_0^1 \theta p(\theta|y)d\theta = \mathrm{E}(\theta|y) = \frac{y+1}{n+2} \qquad (2.6)\end{aligned}$$

である. (2.6) を直接統合してこの結果を再現することは, 演習として残されている。この結果は, 一様事前分布に基づいて, ラプラスの連続法則 (Laplace's law of succession) として知られている。極端な観測値 $y=0$ および $y=n$ において, ラプラスの法則は, それぞれ $\frac{1}{n+2}$, $\frac{n+1}{n+2}$ の確率を予測する.

まとめ 2.1

・ベルヌーイ試行

…成功か失敗がある確率で起こるような試行.

・二項モデル

…ベルヌーイ試行の系列, 二項分布に従うモデル.

$$p(y|\theta) = \mathrm{Bin}(y|n,\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}$$

・二項分布の事前分布

1.Laplace は一様事前分布を用いた.

2. 共役事前分布はベータ分布である.

・ラプラスの連続法則…一様事前分布を用いた二項モデルの事後予測分布は, 尤度関数と連続的な関係にあるということ.

2.2 データと事前情報の間の妥協としての事後分布

ベイズ推論のプロセスは, 事前分布 $p(\theta)$ から事後分布 $p(\theta|y)$ に移行することを含み, いくつかの一般的関係がこれらの 2 つの分布の間に成立すると予想するのは当然である. 例えば, 事後分布はデータからの情報を組み込んでいるため, 前の分布よりも変数の変化が少ないと考えられます. この概念は, 次の式の 2 番目に形式化されています.

$$\mathrm{E}(\theta) = \mathrm{E}(\mathrm{E}(\theta|y)) \tag{2.7}$$

$$\mathrm{var}(\theta) = \mathrm{E}(\mathrm{var}(\theta|y)) + \mathrm{var}(\mathrm{E}(\theta|y)) \tag{2.8}$$

(1.8) と (1.9) の元の (u, v) を (θ, y) で置き換えたものである. 式 (2.7) で表される結果はほとんど驚くべきではない. θ の事前平均は取りうるデータの分布についての全ての取りうる事後平均の平均である. 分散公式 (2.8) はもっと面白い. なぜなら, 可能なデータの分布に対する事後平均の変動に依存する量だけ増加し, それは事後分散が平均において事前平均よりも小さいということを表すからである.

後者の分散が大きければ大きいほど, 次の節の二項モデルと正規モデルについて詳細に説明するように, θ に関する不確かさを減らす可能性が高くなります. 平均と分散の関係は期待値のみを記述し, 特定の状況では, 事後分散は事前分散と類似しているか, またはそれより大きくなる可能性があります (サンプリングモデルと事前分布の間に矛盾または不一致がある場合があります).

一様事前分布に関する二項分布の例では, 事前平均は $\frac{1}{2}$ で, 事前分散は $\frac{1}{12}$ である. 事後平均, $\frac{y+1}{n+2}$ は事前平均とサンプル割合 $\frac{y}{n}$ の間の妥協点である. ここで, 明らかに事前平均はデータサンプルのサイズが大きくなるにつれ小さな役割になっていく. これは, ベイズ推論の一般的な特徴である. 事後分布は事前情報とデータの間の妥協点と表される点に集中し, サンプルサイズが大きくなるつれて, データによって妥協がより大きく支配される.

まとめ 2.2

・事後分布は事前分布と尤度の妥協点

…事後分布は事前情報とデータの間の妥協点と表される点に集中し, サンプルサイズが大きくなるつれて, データによって妥協がより大きく支配される.

2.3 事後推論の要約

事後確率分布はパラメータ θ についての現在の情報全てを含んでいる. 理想的には, 事後分布 $p(\theta|y)$ 全体を報告するかもしれない. 図 2.1 で見るように, 図的に示すことは役に立つ. 3 章で, 多変量の問題で事後分布を示すために輪郭プロットと散布図を用いる. シミュレーションによって実装されるベイズ手法の重要な利点は, 複雑な変換の後でさえ, 事後推論を要約することができる柔軟性である. この利点は, 例を通して最も直接的に理解され, そのうちのいくつかはすぐに提示される.

しかしながら, 多くの実用上の目的に対して, 分布の数値的な要約が望まれる. 位置の通常用いられる要約は分布の平均, 中央値, モードである. 分散は通常, 標準偏差, 四分位範囲, 他の分位数で要約される.

それぞれの要約にはそれぞれ解釈がある. 例えば, データ (とモデル) が与えられた時, 平均はパラメータの事後期待値で, モードは単一の最もありうる値として解釈されるかもしれない. さらに, 我々が見るように, 多くの実用的な推論は, θ に対称化変換を適用することによってしばしば改善される正規近似の使用に依存し, ここでは平均および標準偏差が重要な役割を果たす. このモードは, より複雑な問題の計算戦略では重要です. なぜなら, 平均または中央値よりも計算がしばしば簡単なためである.

事後分布が, 現在の例のベータ分布のように, 閉じた形をしているとき, 事後分布の平, 中央値, 標準偏差がしばしば閉じた形で利用できる. たとえば, 付録 A の分布の結果を適用することによって, (2.3) のベータ分布の平均は $\frac{y+1}{n+2}$ でモードは $\frac{y}{n}$ である. これは, 異なる視点から θ の最大尤度および (最小分散) 偏りのない推定としてよく知られている.

事後分位数と事後区間

点要約に加えて, 事後不確実性を報告することはほとんど常に重要です. 通常のアプローチは, 関心のある推定量の事後分布の分位数を提示することである. もし, 区間要約が望ましい場合, 事後確率の中心区間は, $100(1-\alpha)\%$区間の場合, その値の範囲は事後確率の正確に $100(\alpha/2)\%$である. そのような区間推定値は事後区間と呼ばれる. 二項, 正規分布のような単純なモデルでは, 累積分布関数から直接的に事後区間を計算することができ, これは 2.4 節で例証するように, 人間の性比の例を参照する. 一般に, 間隔は, 1.9 節の終わりに記載されているように, 事後分布からコンピュータシミュレーションを用いて計算することができる.

事後不確実性のわずかに異なる要約が最も高い事後密度領域である. 事後確率の $100(1-\alpha)\%$を含む値の集合であり, 領域内の密度が決して外側の密度よりも低くないという特性も有する. このような領域は, 事後分布が単峰性で対称である場合, 中央事後区間と同一である. 現在の慣行では, 事後 $\alpha/2$ および $1-\alpha/2$ 分位数としての直接的な解釈を有し, 部分的に事後シミュレーションを使用して計算されるため, 中心事後区間が共用される.

図 2.2 は異なる事後要約が更に違って見えるような場合について示している. 95%の最も高い事後密度領域は, 2 つの互いに素な間隔を含むが, 95%の中央区間は分布の中央の 0 確率区間を含む. この状況では, 最も高い事後密度領域はより煩雑であるが, 中心区間よりも多くの情報を伝達する. しかし, おそらく, 任意の単一の間隔でこの二峰性密度を要約しようとしない方が良いでしょう. 事後密度が非常に歪んでいる場合, 中心区間および最高事後密度領域も実質的に異なることがある.

まとめ 2.3

・シミュレーションによって実装されるベイズ手法のメリット

…複雑な変換の後でさえ, 事後推論を要約することができる柔軟性である.

・位置についての要約

…分布の平均, 中央値, モード

・分散についての要約

…標準偏差, 四分位範囲, 他の分位数

・区間推定量

…事後分布において質量が $100(1-\alpha)\%$ 存在する中心の区間などをいう.

2.4 情報事前分布

二項分布の例では, これまで θ の一様事前分布のみを考慮してきました. どのようにしてこの仕様を正当化できるのか, 一般に事前分布を構築する問題にどのようにアプローチするのだろうか?

事前分布に対し与えることができる 2 つの基本的な解釈を考察する. 母集団解釈において, 事前分布は, 現在の関心の θ が描かれた可能なパラメータ値の集団を表す. より主観的な知識解釈では, θ についての知識 (および不確実性) を, あたかもその値が前の分布からのランダムな実現と考えることができるかのように表現しなければならないという原則が導かれる. 新しい産業プロセスにおける失敗の確率を見積もるなどの多くの問題に対して, 仮説的熟考を除いて, 現在の θ が描かれた θ の完全な関連集団は存在しない. 典型的には, 事前分布は θ のすべてのもっともらしい値を含むべきであるが, データに含まれる θ に関する情報が妥当な事前確率仕様を上回ることが多いため, 分布は真の値の周りに現実的に集中する必要はない.

二項分布の例では, θ の一様事前分布は, y(n が与えられた) の事前予測分布が離散集合 $\{0, 1 \cdots, n\}$ と等しく, $n+1$ 個の可能な値に等しい確率を与える. この問題の彼の元々の扱い (2.1 節の歴史的メモに記載されている) では, 一様事前分布に対するベイズの正当性は, この観察に基づいていると思われる. 議論は, 観察可能な量 y と n の観点から完全に表現されているので, 魅力的である. 一様事前密度に関するラプラスの理論的根拠はあまり明確ではなかったが, その後の解釈では θ については何も知られていなければ一様な仕様が適切であると主張するいわゆる不十分な理由の原則 (principle of insufficient reason) が彼によってなされている. 確率分布を割り当てるための一般的なアプローチとして, 不十分な理由の原理の弱点を 2.8 節で議論する.

この時点で, 実質的な情報を反映する事前分布を割り当てる際に生じるいくつかの問題について議論する.

異なる事前分布を持つ二項分布の例

最初に, 特別な場合としての一様分布を含む事前分布のパラメータ族を使用して, 二項モデルをより詳細に追求する. 数学的な便宜のために, 単純な事後密度につながる事前密度族を構築します.

θ の関数として考え, 尤度 (2.1) は

$$p(y|\theta) \propto \theta^a (a-\theta)^b$$

の形の式である. したがって, 事前密度が同じ形式であるならば, 事後密度もこの形式になる. そのような事前分布は

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

としてパラメータ化する. これは, パラメータ α, β に関するベータ分布である. $\theta \sim \text{Beta}(\alpha, \beta)$. $p(\theta), p(y|\theta)$ を比較すれば, 事前密度は事前の $\alpha-1$ 回成功, $\beta-1$ 失敗と等しいということが分かる. 事前分布のパラメータはしばしば超パラメータ (hyperparameter) と呼ばれる. ベータ事前分布は 2 つの超パラメータで特徴づけられる. つまり, 分布の 2 つの特徴, たとえば平均と分散を固定することによって特定の事前分布を指定できます. 583 ページの (A.3) を参照.

今のところ, 合理的な値 α, β を選択できると仮定します. 特定の問題で未知の超パラメータを扱うための適切な方法は, 5 章で説明します. θ の事後密度は,

$$\begin{aligned} p(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta|\alpha+y, \beta+n-y) \end{aligned}$$

事後分布が事前分布と同じパラメトリック形式に従うという性質を共役 (conjugate) と呼びます. ベータ事前分布は, 二項尤度の共役分布族である. 共役分布族は, 事後分布が既知のパラメトリック形式に従うという点で数学的に便利である. 共役分布族に矛盾する情報がある場合は, より現実的で不都合な事前分布を使用する必要があるかもしれません (二項尤度が場合によってはより現実的な可能性に置き換えられる必要があるかもしれません).

ベータ事前分布を伴う二項モデルを続けるために, 母集団からの未来の抽出に対する成功の事後確率として解釈できる θ の事後平均は,

$$\mathrm{E}(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

である. これはサンプル割合 y/n と事前平均 $\alpha/(\alpha+\beta)$ の間にある. 演習 2.5b を参照. 事後分散は

$$\mathrm{var}(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{\mathrm{E}(\theta|y)[1 - \mathrm{E}(\theta|y)]}{(\alpha + \beta + n - 1}$$

である. $y, n-y$ が定数 α, β に対して大きくなるにしたがって, $\mathrm{E}(\theta|y) \simeq y/n$ で $\mathrm{var}(\theta|y) \simeq \frac{1}{n}\frac{y}{n}(1 - \frac{y}{n})$ で, $1/n$ の割合で 0 に近づく. 極限では, 事前分布のパラメータは事後分布で影響を持たない.

実は, 4 章でより詳しく見るように, 確率理論の中心極限定理はベイズの文脈にも使うことができ,

$$\left(\frac{\theta - \mathrm{E}(\theta|y)}{\sqrt{\mathrm{var}(\theta|y)}} \middle| y \right) \to \mathrm{N}(0, 1)$$

この結果はしばしば事後分布の正規分布への近似を正当化するのに使われる. 二項パラメータ θ に対して, もし θ をロジットスケールに変換すると, 正規分布は実際より正確な近似である. すなわち, 確率空間を $[0, 1]$ から $(-\infty, \infty)$ に拡張し, θ 自身の代わりに $\log(\theta/(1-\theta))$ に対して推論を行うことである. これは正規分布に対する近似に対するより良いフィッティングである.

共役事前分布

共役性は正しくは次のように定義される. $\mathcal{F}$ がサンプリング分布 $p(y|\theta)$ のクラスで, $\mathcal{P}$ が θ に対する事前分布のクラスのとき,

$$p(\theta|y) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

ならば, $\mathcal{P}$ は $\mathcal{F}$ に対する共役分布族である. この定義は, 形式的に漠然としている. なぜなら, $\mathcal{P}$ をすべての分布のクラスとして選択すると, どのクラスのサンプリング分布が使用されても, $\mathcal{P}$ は常に共役であるからである. $\mathcal{P}$ を尤度と同じ機能的形態を有するすべての密度の集合とみなすことによって生じる, 自然な共役事前分布に最も興味がある.

共役事前分布は, 計算上の利便性に加えて, 二項の例で見たように, 追加データとして解釈可能であるという実用上の利点を有し, 2.5, 2.6 節の標準および他の標準モデルでも参照する.

非共役事前分布

共役事前分布の使用に関する基本的な正当性は, 尤度に対して標準モデル (二項式および標準など) を使用する場合と似ています. 分析結果をよく理解するのが簡単である. 良好な近似が得られ, 計算が簡単になります. また, 後で共感が不可能な多くの次元を含む, より複雑なモデルの基礎としても役立ちます. これらの理由から, 共役モデルは良い出発点になり得る. 例えば, 共役分布族の混合は, 単純な共役分布が妥当でない場合に有用な場合があります (演習 2.4 を参照).

事後推論の解釈を不透明にし, 計算をより困難にすることがあるが, 非共役事前分布は新しい概念的問題を引き起こさない. 実際には, 複雑なモデルの場合, 共役事前分布は不可能でさえあるかもしれません. 2.4 節と演習 2.10 と 2.11 に非共役計算の例を示します. より広範な非共役の例, バイオアッセイ実験の分析が 3.7 節に示されている.

共役事前分布, 指数型分布族, 十分統計量

本節では, 分布の共役分布族を指数型分布族 (exponential family) の古典概念と十分統計量 (sufficient statistics) に関連づけることで, この節を終了する. これらの概念に精通していない読者は, この例の先に進んでもよい.

指数型分布族に属する確率分布は, 事前分布が自然な共役であるため, この時点で指数型分布族の定義を見直します. この節では完全性のために, データ点 y_i とパラメータ θ を多次元にすることができます. すべての要素が式,

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^{\mathrm{T}}u(y_i)}$$

を持っている場合, クラス $\mathcal{F}$ は指数型分布族である. 要素 $\phi(\theta), u(y_i)$ は一般的に, θ の次元と一致する次元のベクトルである. ベクトル $\phi(\theta)$ は族 $\mathcal{F}$ の自然パラメータと呼ばれる. 独立同分布に従う観測値の系列 $y = (y_1, \cdots, y_n)$ に対応する尤度は

$$p(y|\theta) = \left(\prod_{i=1}^{n} f(y_i)\right) g(\theta)^n \exp\left(\phi(\theta)^{\mathrm{T}} \sum_{i=1}^{n} u(y_i)\right)$$

すべての n, y に対して, これは固定された式を持つ.

$$p(y|\theta) \propto g(\theta)^n e^{\phi(\theta)^{\mathrm{T}} t(y)}, \text{ ただし } t(y) = \sum_{i=1}^{n} u(y_i)$$

θ に対する尤度は $t(y)$ の値を通してのみデータ y に依存するために, 量 $t(y)$ は θ に対する十分統計量と呼ばれる. 十分統計量は, 尤度および事後分布の代数的操作に有用である. もし事前密度が

$$p(\theta) \propto g(\theta)^{\eta} e^{\phi(\theta)^{\mathrm{T}} \nu}$$

であるならば, 事後密度は

$$p(\theta|y) \propto g(\theta)^{\eta+n} e^{\phi(\theta)^{\mathrm{T}} (\nu + t(y))}$$

となる.

これは, 事前分布のこの選択が共役であることを示している. ある種の不規則な場合を除けば, すべての n に対して十分統計量を有する唯一の分布が指数関数的であるため, 一般に, 指数型分布族は自然共役事前分布を有する唯一の分布クラスであることが示されている. 我々は既に n が既知の尤度 $p(y|\theta, n) = \mathrm{Bin}(y|n, \theta)$ に対して, θ 上の共役事前分布はベータ分布である二項分布について議論した. 二項分布が自然パラメータ $\mathrm{logit}(\theta)$ を持つ指数型分布族であることを示すための練習として残しておきます.

例. 前置胎盤のとき女の子が出生する確率

性比に影響を及ぼす要因の具体例として, 正常な膣分娩から胎児を塞ぐ, 胎盤が子宮内に埋め込まれている妊娠の異常な状態である母体状態である胎盤前庭を考慮する. ドイツの前置胎盤の出生の性に関する初期の研究では, 合計 980 の出生のうち, 437 人が女性であった. 胎盤の出生前の出生集団における女性の出生の割合が 0.485 未満であるという主張は, どの程度の証拠を提供しているのか? 一般集団における女性の出生の割合?

一様事前分布を用いる分析

少女の出生の確率に対し一様事前分布のもとで, 事後分布は Beta(438, 544) である. 事後分布の正確な要約は, ベータ分布の特性から得ることができる (付録 A). θ の事後平均は 0.446 であり, 事後標準偏差は 0.016 である. ベータ密度の数値積分を使用して正確な事後分位数を得ることができます. ベータ密度は, 実際にはコンピュータ関数呼び出しによって実行されます. 中央値は 0.446 であり, 中央の 95%後方間隔は [0.415, 0.477] である. この 95%事後区間は, 計算された事後平均と標準偏差との正規近似を用いて得られるであろう間隔を小数点以下 3 桁に一致させる. 事後分布の近似的な正規性についてのさらなる議論は 4 章で与えられる.

多くの場合, 事後密度関数の計算を直接実行することは不可能です. そのような場合, 事後分布からのシミュレーションを使用して推論を得ることは特に有用である. 図 2.3 の最初のヒストグラムは, Beta(438, 544) 事後分布からの 1000 個の描画の分布を示しています. 1000 個の抽出の 25 番目と 976 番目を取ることによって得られた 95%の事後区間の推定値は [0.415, 0.476] であり, 事後分布からの 1000 個の抽出の中央値は 0.446 である. 1000 個の抽出のサンプル平均と標準偏差は 0.445 と 0.016 で、正確な結果とほぼ同じです.

95%の事後間隔に対する通常の近似は, $[0.445 \pm 1.96 \cdot 0.016]$=[0.414, 0.476] である. 大きなサンプルと, θ の分布が 0 と 1 から離れて集中するという事実のために, この例では正常な近似がうまくいく.

すでに述べたように, 割合を見積もるとき, 通常の近似は, パラメータ空間を単位区間から実線に変換する logit 変換 $\log(\frac{\theta}{1-\theta})$ に適用することで一般的に改善されます. 図 2.3 の 2 番目のヒストグラムは, 変換された描画の分布を示しています. 1000 回の抽出に基づくロジットスケールの推定後平均および標準偏差は, -0.220 および 0.065 である. θ の 95%事後区間に対する正規近似は, logit スケール $[-0.220 \pm 1.96 \cdot 0.065]$ で 95%間隔を反転することによって得られ, 元のスケールでは [0.414, 0.477] となる. ロジットスケールを使用することによる改善は, サンプルサイズが小さい場合, または θ の分布が 0 または 1 に近い値を含む場合に最も顕著である. 実際のデータ分析では, 適用されるコンテキストを念頭に置いておくことが重要です. この例の関心のあるパラメータは, 伝統的に、男女の出生率である性別比率 $(1-\theta)/\theta$ として表されています. 比率の事後分布は, 第 3 のヒストグラムに示されている. 性比の事後中央値は 1.24 であり, 事後区間の 95%は [1.10, 1.41] である. 事後分布は, 通常のヨーロッパ人種比 1.06 をはるかに上回る値に集中しており, 前置胎盤での女性の出生の確率は一般の人口よりも低いことを意味している.

異なる共役事前分布を用いる分析

提案された事前分布に対する θ に関する事後推論の感度は, 表 2.1 に示されている. 第 1 行は均一な事前分布 $\alpha = 1, \beta = 1$ に対応し, 表の後の行は, 一般集団における女性の出生の割合である 0.485 付近にますます集中する事前分布を使用する. 最初の列は θ の事前平均を示し, 2 番目の列は $\alpha + \beta$ によって測定された事前情報の量を示します. ある意味では, $\alpha + \beta - 2$ は以前の観測数に等しいことを想起されたい. 大きなサンプルに基づく事後推論は, 事前分布に特に敏感ではない. 前回の分布に 100 または 200 の出生に相当する情報が含まれている表の下部にのみ, 事後区間が前の分布に向けて目立つように引かれていても, 95%の事後区間が依然として前の平均を除外しています.

異なる非共役事前分布を用いる分析

この問題に対する共役ベータ分布族の代替として, 0.485 を中心とする事前分布を好むかもしれないが, 真理が遠い可能性を認めるためには、この値から遠くに平坦である. 図 2.4a の区分的線形事前密度は, この形式の事前分布の例です. 確率質量の 40%が区間 [0.385,0.585] の外側にある. この事前分布は平均 0.493 および標準偏差 0.21 を有し, $\alpha + \beta = 5$ のベータ分布の標準偏差と同様である. 正規化されていない事後分布は, 各点における前の密度と二項尤度を乗算することによって, θ 値 $(0.000, 0.001, \cdots, 1.000)$ のグリッドで得られる. 事後シミュレーションは, θ 値の離散グリッド上の分布を正規化することによって得ることができる. 図 2.4b は, 離散的な事後分布からの 1000 回の抽出のヒストグラムです. 事後中央値は 0.448 であり, 中央事後区間の 95%は $[0.419, 0.480]$ である. 事前分布はデータに圧倒されているため, これらの結果はベータ分布に基づいて表 2.1 の結果と一致します. グリッドアプローチをとるには, あまりにも粗いグリッドを避け, 事後質量のかなりの部分を歪ませることが重要です.

まとめ 2.4

・不十分な理由の原則

…Laplace が二項モデルの事前分布として, 一様分布を用いたように, 何も知らない場合は一様分布を用いるべきであるという主張.

・超パラメータ

…事前分布のパラメータのこと.

・共役

…事後分布が事前分布と同じパラメトリック形式に従うという性質をいう. その分布族のことを共役分布族という.

・二項分布の共役事前分布

…Beta(α, β) で, 超パラメータは α 回の成功, β 回の失敗と考えられる.

・非共役事前分布

…事後推論の解釈を不透明にし, 計算を困難にすることがあるが, 概念的に問題はない. 複雑なモデルの場合は使えない.

・指数型分布族

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^{\mathrm{T}}u(y_i)}$$

2.5 既知の分散で正規分布の平均を推定する

正規分布は, ほとんどの統計的モデリングにとって基本的なものである. 中央極限定理は, 解析的に都合のよい実際の可能性の近似として, 多くの統計的問題における通常の尤度を用いることを正当化するのに役立つ. また, 後の章で見るように, 正規分布がそれ自体良好なモデル適合を提供しない場合であっても, t 分布または有限混合分布を含むより複雑なモデルの構成要素として有用であり得る. 今のところ, 通常のモデルが適切であると仮定してベイズの結果を調べるだけです. 最初に単一のデータポイントについて結果を導き出し, 次に多くのデータポイントの一般的なケースについて結果を導き出します.

1つのデータ点に対する尤度

最も簡単な最初の場合として, 平均 θ と分散 σ^2 でパラメータ化される正規分布に従う単一のスカラー観測値 y について考える. ここで, この初期の発展のために, σ^2 が既知であると仮定する. サンプル分布は

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$$

である.

共役事前分布と共役事後分布

θ の関数として考え, 尤度は θ の二次形式の指数である. なので, 共役事前密度の族は

$$p(\theta) = e^{A\theta^2+B\theta+C}$$

である. この族を

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

としてパラメータ化する. すなわち, 超パラメータ μ_0, σ_0^2 に関して, $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$ である. この予備的開発ではいつものように, 超パラメータが既知であると仮定する.

共役事前密度は θ に対する事後分布は二次形式の指数つまり正規分布であるということを意味するが, その具体的な形を明らかにするにはいくつか計算が必要である. 事後密度において, θ を除く全ての変数は定数とみなされ, 条件付き密度を与える.

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)$$

指数を展開し, θ の二乗のを集め完成させれば (詳細については演習 2.14(a) を参照),

$$p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) \tag{2.9}$$

すなわち, $\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2)$ である. ここで,

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \text{ and } \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2} \tag{2.10}$$

事前分布と事後分布の精度

正規分布を操作する際, 分散の逆数が顕著な役割を果たし, 精度と呼ばれる. 上記の計算は, 通常のデータおよび通常の事前分布 (それぞれが既知の精度を有する) の場合, 事後精度が事前精度とデータの精度の和に等しいことを示している.

事後平均 μ_1 の式の解釈の方法はいくつも存在する. (2.10) では, 事後平均は事前平均と観測値 y の, 精度に比例する重みに対する重み付き平均として表現される. あるいは, 観測された y に対して調整された事前平均として μ_1 を

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

と表すことができ, もしくは, 事前平均に縮小したデータとして,

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma^2 + \tau_0^2}$$

と表すことができる. 各式は, 事前平均と観測値との間の妥協点として事後平均を表す.

極端な場合, 事後平均は事前平均または観測データに等しい.

$$\mu_1 = \mu_0 \text{ if } y = \mu_0 \text{ or } \tau_0^2 = 0$$
$$\mu_1 = y \text{ if } y = \mu_0 \text{ or } \sigma^2 = 0$$

もし $\tau_0^2 = 0$ であるならば, 事前分布はデータよりも正確で, 従って事後分布及び事前分布は同一であり, 値 μ_0 に集中する. $\sigma^2 = 0$ の場合, データは完全に正確であり, 事後分布は観測値 y に集中する. $y = \mu_0$ であれば, 事前平均とデータ平均は一致し, 事後平均もこの点とならなければならない.

事後予測分布

未来の観測値 $\tilde{y}$ の事後予測分布は (1.4) を用いて, 直接積分することで計算できる,

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta \propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)d\theta$$

1 つ目は, θ が与えられたもとで, 未来の観測値 $\tilde{y}$ の分布が過去のデータ y に依存しないために成り立つ.

$\tilde{y}$ の分布はに変数正規分布の特性を用いることでより簡単に決定できる. 積分内の席は $(\tilde{y}, \theta)$ の二次関数の指数である. 従って, $\tilde{y}, \theta$ は結合事後正規分布を持ち, $\tilde{y}$ の周辺事後分布は正規分布である..

(2.7), (2.8) の等式と事後分布からの知識, $\mathrm{E}(\tilde{y}|\theta) = \theta$, $\mathrm{var}(\tilde{y}|\theta) = \sigma^2$ を用いることで, 事後予測分布の平均, 分散を決定できる.

$$\mathrm{E}(\tilde{y}|y) = \mathrm{E}(\mathrm{E}(\tilde{y}|\theta, y)|y) = \mathrm{E}(\theta|y) = \mu_1$$

$$\mathrm{var}(\tilde{y}|y) = \mathrm{E}(\mathrm{var}(\tilde{y}|\theta, y)|y) + \mathrm{var}(\mathrm{E}(\tilde{y}|\theta, y)|y) = \mathrm{E}(\sigma^2|y) + \mathrm{var}(\theta|y) = \sigma^2 + \tau_1^2$$

よって, $\tilde{y}$ の事後予測分布は θ の事後平均と同じ平均, 分散の 2 つの要素, モデルからの予測分散 σ^2 と θ の事後不確実性に対する分散 τ_1^2 をもつ.

複数の観測値に関する正規モデル

単一観測値に関する正規モデルの開発は, 独立同分布に従う観測値 $y = (y_1, \cdots, y_n)$ のサンプルが利用できるより現実的な状況に簡単に拡張できる. 依然と同じように, 事後密度は

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta)p(y|\theta) = p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\right)
\end{aligned}
$$

この式の代数的簡素化 (演習 2.14(b) で説明されているように単一観察ケースで使用されているのと同様の線に沿って）は, 事後分布はサンプル平均 $\bar{y}$ だけに依存することを示しています. つまり, このモデルでは $\bar{y}$ は十分統計量です.

実は, $\bar{y}|\theta, \sigma^2 \sim \mathrm{N}(\theta, \sigma^2/n)$ なので, 単一正規観測値で導かれる結果は適用すれば,

$$p(\theta|y_1, \cdots, y_n) = p(\theta|\bar{y}) = \mathrm{N}(\theta|\mu_n, \tau_n^2) \tag{2.11}$$

を得られる. ここで,

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \; and \; \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \tag{2.12}$$

である. なお, 1 回ずつ次の事前分布として事後分布を用いることで, データ $y_1, \cdots, y_n$ に対して情報を追加することで, 同じ結果が得られる.

事後平均と事後分散に対する式の中で, 事前精度 $1/\tau_0^2$ とデータ精度 n/σ^2 は等しい役割をし, n が大きいならば, 事後分布は $\sigma^2, \bar{y}$ で大きく決定される. たとえば, $\tau_0^2 = \sigma^2$ ならば, μ_0 の値に関して事前分布は追加の観測値として同じ重みをもつ. さらに, 定数 n に対して $\tau_0 \to \infty$, もしくは定数 τ_0^2 に対して $n \to \infty$ とすれば,

$$p(\theta|y) \simeq \mathrm{N}(\theta|\bar{y}, \sigma^2/n) \tag{2.13}$$

を得る. これは, 実用で, 尤度が実際的であり, 事前分布の信用が比較的 θ の範囲においてあいまいなとき, よい近似となる.

まとめ 2.5

・正規モデル (単一観測に対する, ただし, 平均未知, 分散既知)

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} (= \mathrm{N}(\theta, \sigma^2))$$

・正規モデルの平均 θ に対する共役事前分布.

…$\mathrm{N}(\mu_0, \sigma_0^2)$ である. また, 事後分布は正規分布となる.

$$p(\theta|y) = \mathrm{N}(\mu_1, \tau_1^2),\ \mu_1 = \frac{1/\tau_0^2\mu_0 + 1/\sigma^2 y}{1/\tau_0^2 + 1/\sigma^2} \text{ and } \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

・事後平均についての解釈

…事後平均は事前精度とデータ精度を重みとする, 事前平均とデータ平均の重み付き平均である.

・事後予測分布

$$p(\tilde{y}|y) = \mathrm{N}(\mu_1, \sigma^2 + \tau_1^2)$$

と直接積分から求めることができ, 平均は同じであるが分散は大きくなる.

・正規モデル (複数観測に対する, ただし, 平均未知, 分散既知)

…idd に従う観測値 $y = (y_1, \ldots, y_n)$ についても考えるが, ほとんどアナロジーと一致する. $\bar{y}$ が十分統計量であるので, $\bar{y}|\theta, \sigma^2 \sim \mathrm{N}(\theta, \sigma^2/n)$ なので, 単一正規観測値で導かれる結果は適用すれば,

$$p(\theta|y_1, \ldots, y_n) = p(\theta|\bar{y}) = \mathrm{N}(\theta|\mu_n, \tau_n^2),\ \mu_n = \frac{1/\tau_0^2\mu_0 + n/\sigma^2 y}{1/\tau_0^2 + n/\sigma^2} \text{ and } \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

を得られる.

・事後平均についての解釈

…n が大きくなるにしたがって, データ精度の重みが大きくなることが分かる. また, 分散は小さくなっていく.

・事後予測分布

…分散が小さくなっていき, より良い予測が可能になるという点で, 感覚と一致する

2.6 他の標準的な単一パラメータモデル

一般に, 事後密度 $p(\theta|y)$ は閉形式を持たないことを想起する. 正規化定数 $p(y)$ は, 積分 (1.3) のために計算することがしばしば困難です. 正式なベイズ分析の多くは, 閉じた形式が利用可能である状況に集中します. そのようなモデルは時には非現実的ですが, より現実的なモデルを構築するためには, その分析が役立つ出発点となることがよくあります.

二項, 正規, ポアソン, 指数分布の標準分布は, 単純確率モデルから自然に導かれます. すでに議論したように, 二項分布は交換可能な結果を数えることによって動機づけられ, 正規分布は多くの交換可能な独立項の和である確率変数に適用されます. また, 正のデータの対数に正規分布を適用する機会がありますが, 多くの独立した乗法因子の積としてモデル化された観測に自然に適用されます. ポアソン分布および指数分布は, すべての時間間隔で交換可能に発生するとモデル化されたイベントのカウント数および待ち時間としてそれぞれ発生します. すなわち, 時間的に独立して, 発生率が一定である. これらの基本的な分布を組み合わせることで, より複雑な結果を得るための現実的な確率モデルを一般的に構築する. たとえば, 22.2 節では, 私たちは心理学的実験における精神分裂病患者の反応時間を, 対数スケールでの正規分布の二項混合としてモデル化する.

これらの標準モデルには, 共役事前分布族があり, それを議論する.

既知の平均と未知の分散をもつ正規分布

既知の平均と未知の分散をもつ正規分布は, 必ずしも直接適用される値ではないが, より複雑で役に立つモデル, 最もすぐに見るのは未知の平均と分散をもつ正規分布の基礎として重要な例である. これは, 3.2 節でカバーされる. 加えて, 既知の平均と未知の分散をもつ正規分布はスケールパラメータの推定の導入的な例を与える.

既知の θ と未知の σ^2 に関する $p(y|\theta,\sigma^2) = \mathrm{N}(y|\theta,\sigma^2)$ に対して, n 個の独立同分布に従う観測値のベクトル y に対する尤度は

$$\begin{aligned} p(y|\sigma^2) &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) \\ &= (\sigma^2)^{-n/2}\exp\left(-\frac{n}{2\sigma^2}v\right) \end{aligned}$$

十分統計量と対応する共役事前密度, 逆-ガンマ分布は

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2,\ p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}$$

である. ここで (α,β) は超パラメータである.

便利なパラメータ化は, スケール σ_0^2 と自由度 ν_0 を持つスケーリングされた逆 χ^2 分布である. (付録 A を参照). すなわち, σ^2 の事前分布は $\sigma_0^2\nu_0/X$ の分布となる. ここで, X は $\chi^2_{\nu_0}$ 確率変数である. 便利であるが標準的でない表記 $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ を用いる.

結果として出てくる σ^2 に対する事後分布は,

$$p(\sigma^2|y) \propto p(\sigma^2)p(y|\sigma^2) \propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\nu_0/2+1} \exp\left(-\frac{\nu_0\sigma^2}{2\sigma^2}\right) \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}\frac{v}{\sigma^2}\right)$$

$$\propto (\sigma^2)^{-((n+\nu_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + nv)\right)$$

となる. こうして,

$$\sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}\right)$$

これは, 事前分布とデータのスケールの自由度重み付き平均および事前分布の自由度およびデータの自由度の合計に等しい自由度にスケーリングされた逆-χ^2 分布である. 事前分布は, 平均 2 乗偏差 σ_0^2 を有する ν_0 個の観測に相当する情報を提供するものと考えることができる.

ポアソンモデル

ポアソン分布はカウントの形をとるデータの研究で自然に表れる. 例えば, 主な適用分野は疫学であり, 疾病の発生率が研究されている.

もしデータ点が割合 θ に関するポアソン分布に従うならば, 単一観測値 y の確率分布は

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \text{ for } y = 0, 1, \cdots$$

で, 独立同一分布に従う観測値のベクトル $y = (y_1, \cdots, y_n)$ の尤度は,

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \propto \theta^{t(y)} e^{-n\theta}$$

である. ここで, $t(y) = \sum_{i=1}^{n} y_i$ は十分統計量である. 尤度は指数型分布族の式で, 自然パラメータを $\phi(\theta) = \log \theta$ とし,

$$p(y|\theta) \propto e^{-n\theta} e^{t(y) \log \theta}$$

と書き換えられる.

また, 自然な共役事前分布は, 超パラメータ (η, ν) で

$$p(\theta) \propto (e^{-\theta})^{\eta} e^{\nu \log \theta}$$

となる. この議論を別の方法で表現すると, 尤度は $\theta^a e^{-b\theta}$ の形式であり, 共役事前密度は $p(\theta) \propto \theta^A e^{-B\theta}$ の形式でなければならない. より便利なパラメータ化では,

$$p(\theta) \propto e^{-\beta\theta} \theta^{\alpha-1}$$

である. これは, パラメータ α, β のガンマ分布である.Gamma(α, β). 付録 A を参照. $p(y|\theta), p(\theta)$ を比較すると, 事前の密度はある意味では, β 個の事前行観測における $\alpha - 1$ の総数に等しいことが分かる. 共役事前分布に関して, 事後分布は

$$\theta|y \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$$

となる.

負の二項分布

共役分布族に関して, 事前分布, 事後分布の既知の式は, 公式

$$p(y)_= \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$

を用いることで, 周辺分布 $p(y)$ を見つけるのに使える. 例えば, 単一観測値 y に対するポアソンモデルは, 事前予測分布は

$$\begin{aligned} p(y) &= \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,1+\beta)} \\ &= \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}} \end{aligned}$$

であり, これは,

$$p(y) = \left(\begin{array}{c} \alpha+y-1 \\ y \end{array} \right) \left(\frac{\beta}{\beta+1} \right)^{\alpha} \left(\frac{1}{\beta+1} \right)^{y}$$

を導き, さらにこれは負の二項密度として知られている.

$$y \sim \text{Neg-bin}(\alpha,\beta)$$

17.2 節でポアソン分布のロバスト代替として負の二項分布に戻る.

割合と暴露に関してパラメータ化されたポアソンモデル

多くの応用で, データ点 $y_1, \cdots, y_n$ に対して

$$y_i \sim \text{Poisson}(x_i\theta) \tag{2.14}$$

とポアソンモデルを拡張することは役に立つ. ここで, 値 x_i は既知正の値をとる説明変数で θ は興味のある未知のパラメータである. 疫学では, θ はしばしば割合と, x_i は i 番目の暴露呼ばれる. このモデルは y_i において交換可能でないが, $(x, y)_i$ において交換可能である. 拡張されたポアソンモデルでの θ に対する尤度は (θ に依存しないパラメータは無視し)

$$p(y|\theta) \propto \theta^{(\sum_{i=1}^n y_i)} e^{-(\sum_{i=1}^n x_i)\theta}$$

で, θ に対するガンマ分布が共役である. 事前分布

$$\theta \sim \text{Gamma}(\alpha, \beta)$$

に関して, 結果となる事後分布は

$$\theta|y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n x_i\right) \tag{2.15}$$

である.

ポアソンデータから割合の推定: 理想化された例

米国の都市で 1 年間死亡原因が詳細に検討されているとします. 人口 20 万人のうち 3 人が喘息で亡くなり, 年間 10 万人あたり 1.5 人の都市で推定喘息死亡率が明らかになった. ポアソンサンプリングモデルは, この形式の疫学データによく使用されます. ポアソンモデルは, すべての小さな曝露間隔の間の交換可能性の仮定から導かれる. ポアソンモデルでは, 1 年間の 20 万人の都市における死亡数 y のサンプリング分布は, Poisson(2.0θ) として表すことができます. ここで, θ は, 本市の真の基礎的な長期喘息死亡率を表します (年間 10 万人あたりの症例で測定). 上記の表記法では, $y = 3$ は $x = 2.0$ の露出 (θ は 100,000 人の単位で定義されるため) と未知の率 θ を持つ単一の観測値です. 世界中の喘息死亡率に関する知識を使って, θ の事前分布を作成し, $y = 3$ のデータをその事前分布と組み合わせて事後分布を得ることができます.

事前分布の設定

θ の合理的な事前分布はなんだろうか? 世界中の喘息死亡率のレビューから, 10 万人あたり 1.5 人を超える死亡率は西洋諸国では稀で, 典型的な喘息死亡率は 10 万人あたり 0.6 人程度であることが示唆されています. この問題のために以前の共役分布族であったガンマ分布の特性を試行錯誤してみると, この例の Gamma(3.0, 5.0) 密度は, この都市と他の都市と今年と他の年の間の交換可能性を仮定するならば, この例の喘息死亡率のもっともらしい事前密度を与えることがわかります. この事前分布の平均は 0.6(モードは 0.4) であり, 密度の質量の 97.5%は 1.44 未満である. 実際には, 事前平均を指定することにより, 2 つのガンマパラメータの比が設定され, 形状パラメータは, 分布の尾部に関する事前知識と一致するように試行錯誤によって変更することができる.

事後分布

（2.15）の結果は, Gamma(α, β) 事前分布の θ の事後分布が Gamma($\alpha + y, \beta + x$) であることを示している. 前の分布および記述されたデータでは, θ の事後分布は Gamma(6.0, 7.0) であり, これは平均 0.86 であり、実質的な収縮は前の分布に向かって生じている. θ の事後分布から 1000 回のヒストグラムを図 2.5a に示す。例えば, 私たちの街の喘息による長期死亡率がガンマ事後密度から計算された年 10 万人あたり 1.0 を超える事後確率は 0.30 である.

指数モデル

指数分布は一般的に待ち時間モデルやしばしば時間スケールの連続で正の実数値確率変数などに使われる. パラメータ θ が与えられた, 結果 y のサンプリング分布は

$$p(y|\theta) = \theta \exp(-y\theta), \text{ for } y > 0$$

で, $\theta = 1/\mathrm{R}(y|\theta)$ は割合と呼ばれる. 数学的に, 指数分布はパラメータを $(\alpha, \beta) = (1, \theta)$ としたガンマ分布の特別な場合である. この場合では, しかしながら, ポアソン分布の例と同じようにパラメータ θ の事前分布として使うのではなく, 結果 y に対するサンプリング分布として使われている.

指数分布には, 生存または生涯データの自然なモデルとなる無記憶性の特性がある. オブジェクトが追加の長さの時間 t で存続する確率は, この時点までに経過した時間とは無関係です. 任意の s, t に対して
$\Pr(y > t + s|y > s, \theta) = \Pr(y > t|\theta)$.

指数分布のパラメータ θ に対する共役事前分布は, ポアソン分布の平均と同じく, $\text{Gamma}(\theta|\alpha, \beta)$ で, 対応する事後分布は $\text{Gamma}(\theta|\alpha+1, \beta+y)$ である. 定数の割合 θ で n 個の独立同分布に従う観測値 $y = (y_1, \cdots, y_n)$ のサンプリング分布は

$$p(y|\theta) = \theta^n \exp(-n\bar{y}\theta), \text{ for } \bar{y} \geq 0$$

これは, θ の尤度として見られるとき, $\text{Gamma}(n+1, n\bar{y})$ に比例する. これより, θ に対しする $\text{Gamma}(\alpha, \beta)$ 事前分布は総待ち時間 β で $\alpha - 1$ 個の指数観測値としてみることができる. (演習 2.19 を参照).

まとめ 2.6

・二項分布

…交換可能な結果を数えるとき用いられる.

・正規分布

…交換可能な独立項の和である確率変数に適用される. また, 正のデータの対数に正規分布を適用する機会がある.

・ポアソン分布, 指数分布

…すべての時間間隔で交換可能に発生するとモデル化されたイベントのカウント数および待ち時間としてそれぞれ発生します.

・正規モデル (ただし, 平均既知, 分散未知)

…十分統計量はサンプル分散で, 共役事前分布は逆-ガンマ分布である.

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2,\ p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}$$

便利なパラメータ化逆-χ^2 分布を用いると, 事前分布, 事後分布は

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2),\ \sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}\right)$$

・ポアソンモデル (ただし, 割合未知)

…割合 θ のポアソン分布に従うなら, 単一観測値 y の確率分布は

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \text{ for } y = 0, 1, \cdots$$

で, 独立同一分布に従う観測値のベクトル $y = (y_1, \cdots, y_n)$ の尤度は,

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta}$$

共役事前分布はガンマ分布, $\text{Gamma}(\alpha, \beta)$ で事後分布は

$$\theta|y \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$$

単一観測値 y のポアソンモデルの事前予測分布は負の二項分布

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)} = \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,1+\beta)} = \frac{\Gamma(\alpha+y)\beta^\alpha}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}}$$

$$= \begin{pmatrix} \alpha + y - 1 \\ y \end{pmatrix} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^y = \text{Neg-bin}(\alpha, \beta)$$

・ポアソンモデル (割合と暴露のモデル)

…データ点 $y_1, \ldots, y_n$ に対して,

$$y_i \sim \text{Poisson}(x_i\theta)$$

とするモデルがある. x_i は暴露で θ は割合である. θ に対する共役事前分布はガンマ分布であり, ともに事後分布は

$$\theta \sim \text{Gamma}(\alpha, \beta),\ \theta|y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} x_i\right)$$

・指数モデル

…一般的に待ち時間モデルやしばしば時間スケールの連続で正の実数値確率変数などに使われる. パラメータ θ が与えられた, 結果 y のサンプリング分布は

$$p(y|\theta) = \theta \exp(-y\theta),\ \text{for } y > 0$$

無記憶性をもつ. 任意の s, t に対し $\Pr(y > t + s|y > s, \theta) = \Pr(y > t|\theta)$.
ポアソン分布と同じく, 共役事前分布は $\text{Gamma}(\theta|\alpha, \beta)$ で, 対応する事後分布は $\text{Gamma}(\theta|\alpha + 1, \beta + y)$ である.

2.7 例: がんの割合に対する情報事前分布

2.4 節の終わりで固定量のデータが与えられた時の推論における事前分布の影響を考えた. ここでは, 対照的に, 共通の事前分布を持つことなるデータに基づく推論の大きな集合について考える. 事前分布の役割を示すことに加えて, この例は 5 章で戻る階層モデルの導入も行う.

地図のややこしいパターン

図 2.6 は, 1980 年代の腎がん死亡率が最も高い米国の郡を示しています. 地図の中で最も目立つパターンは, 国の真ん中にある大平原の郡の多くが, 海岸近くの郡は比較的少数です.

地図を表示すると, 人々は多くの理論を出して, グレートプレーンズの不均等な陰影を説明します. おそらく空気や水が汚染されている, あるいは人々が治療を受けていない傾向があり, がんが治療に遅すぎるとか, あるいは食事が不健康であることがあります. これらの推測はすべて真実かもしれないが, 実際には図 2.6 のパターンを説明する必要はない. これを見るには, 腎臓がん死亡率が最も低い郡の 10%をプロットする図 2.7 を見る. これらはまた, ほとんどが国の真中にある. だから, なぜこれらの地域が最低料金と最高料金を持っているのかを説明する必要がある.

問題はサンプルサイズです. 人口 1000 人の郡を考えてみましょう. 腎臓がんはまれな疾患であり, 10 年間で 1000 人の郡が腎臓がんの死亡率をゼロにしている可能性が高いため, 図 2.7 に影を付ける. しかし, この郡では 10 年間で腎臓がんの死亡者が 1 人になる可能性があります. そうであれば, それは年間 10,000 人に 1 人の割合であり, それは図 2.6 で網掛けされるように上位 10%に入るのに十分高い. グレートプレーンズには人口が少ないので, どちらの地図でも過度に表示されています. これらの地図から, がん率が特に高いという証拠はありません.

がんの死亡率のベイズ推論

そのままの割合のマップにおける誤解を招くパターンは, 真の潜在的な割合を推定するためのモデルに基づくのアプローチが有用であることを示唆している. 特に, モデル

$$y_i \sim \text{Poisson}(10 n_j \theta_j) \tag{2.16}$$

を用いて各郡 j での潜在的ながんの死亡率を推定することは自然である. ここで, y_j は 1980-1989 年の郡 j における腎臓がんの死亡数であり, n_j は郡の人口であり, θ_j は 1 人当たりの死亡者数の潜在的な割合である. この表記法では, 図 2.6 と図 2.7 のマップは, もともとの割合 $\frac{y_j}{10 n_j}$ をプロットしています. (ここでは, 年齢標準化を無視していますが, これを可能にするモデルの一般化は可能である.)

このモデルは θ_j が郡によって異なるという点で (2.14) とは異なり, (2.16) は米国の各郡の別個のモデルである. (2.16) の添え字 j(i ではなく) を使用する. 個々のパラメータは, それぞれがそれ自身のデータから推定される. 郡のうちの 1 つだけの推論を行っていたら, $y \sim \text{Poisson}(10n\theta)$ と書くだけである.

ベイズ推論を行うために, 未知の割合 θ_j に対する事前分布が必要となる. のちに議論するように, パラメータ $\alpha = 20, \beta = 430,000$ に関するガンマ分布は U.S. の郡の潜在的な肝臓がんの死亡割合に対する合理的な事前分布である. この事前分布は $\frac{\alpha}{\beta} = 4.65 \times 10^{-5}$ の平均と標準分散 $\frac{\sqrt{\alpha}}{\beta} = 1.04 \times 10^{-5}$ を持つ.

θ_j の事後分布は,

$$\theta_j | y_j \sim \text{Gamma}(20 + y_j, 430,000 + 10n_j)$$

で, これは平均と分散が

$$\text{E}(\theta_j | y_j) = \frac{20 + y_j}{430,000 + 10n_j}, \ \text{var}(\theta_j | y_j) = \frac{20 + y_j}{(430,000 + 10n_j)^2}$$

である. 事後平均はもともとの割合 $\frac{y}{10n_j}$ と事前平均 $\frac{\alpha}{\beta} = 4.65 \times 10^{-5}$ の重み付き平均としてみることができる. (似た計算については演習 2.5 を参照)

ローカルデータと事前分布の相対的な重要性

小さい郡に対する推論

事前情報とデータの相対的な重みづけは母集団のサイズ n_j に依存する. 例えば, $n_j = 1000$ の小さな郡を考えればよい.

・この郡に対して, $y_j = 0$ ならば, もともとの死亡率は 0 であるが事後平均は $\frac{20}{440,000} = 4.55 \times 10^{-5}$ となる.

・$y_j = 1$ ならば, もともとの死亡率は 10 年で 1000 人に 1 人もしくは 1 年人につき 10^{-4} であるが, 事後平均は $\frac{21}{440,000} = 4.77 \times 10^{-5}$ である.

・$y_i = 2$ ならば, もともとの死亡率は 10 年で 1000 人に 1 人もしくは 1 年人につき 2×10^{-4} で極めて高いが, 事後平均は $\frac{22}{440,000} = 5.00 \times 10^{-5}$ である.

このような小さな集団サイズでは, データは事前分布によって支配されています.

しかし, 先験的に, この郡にとって, y_j は 0,1,2 などとどれくらい同じであろうか. $nj = 1000$? これは予測分布, y_j の周辺分布, θ_j の事前分布にわたって平均することによって決定される. 2.6 節で説明したように, ガンマ事前分布を持つポアソンモデルは, 負の二項予測分布を持っています.

$$y_i \sim \text{Neg-bin}\left(\alpha, \frac{\beta}{10n_j}\right)$$

y_j の予測分布を次のように直接シミュレートする方が簡単かもしれません. (1) Gamma$(20, 430, 000)$ 分布から θ_j の値を 500(例えば) 引き出す. (2) これらのそれぞれについて, パラメータ $10{,}000\theta_j$ を有するポアソン分布から 1 つの値 y_j を引き出す. このようにして生成された y_j の 500 回のシミュレーションのうち, 319 個は 0 であり, 141 個は 1 であり, 33 個は 2 であり, 5 個は 3 であった.

大きな群に対する推論

今, n_j =100 万の大きな郡を考えてみましょう. 10 年の間に何人のがんの死亡が見込まれますか? ここでもまた, Gamma(20, 430, 000) および Poisson($107\theta_j$) 分布を使用して, 予測分布から 500 個の値 y_j をシミュレートすることができます. これを行うと, 473 の中央値と [393,545] の 50%の区間が見つかりました。そのような郡における生死の率は, 3.93×10^{-5} と 5.45×10^{-5} との間になる可能性が高いかそうでないかになります.

ベイズ的に推定されたベイズ調整された死亡率はどうですか? この大規模な郡では, データが事前分布を支配しています.

異なるサイズの群の比較

ポアソンモデル (2.16) では, $\frac{y_j}{10n_j}$ の分散は曝露パラメータ n_j に反比例するため, 郡 j のサンプルサイズとみなすことができます. 図 2.8 は, もともとの腎臓がんの死亡率が人口によってどのように異なるかを示している. 非常に高い割合と非常に低い割合はすべて, 人口が少ない国にあります. 比較すると, 図 2.9a は, ベイズ推定割合がはるかに小さいことを示しています. 最後に, 図 2.9b は, 郡のサンプル (1 つのプロットにすべて 3071 を表示するのが難しいため選択) の 50%区間の推定値を表示します. 小規模の郡では情報が少ないため, 広範囲の事後区間があります.

事前分布の構築

ここで元に戻って元のレートの Gamma(20, 430, 000) の事前配分を得た場所について議論します. モデルを導入するときに議論したように, 我々は数学的な便宜のためにガンマ分布を選んだ. ここで, 観測された癌死亡率 $\frac{y_j}{10n_j}$ の分布と一致するように, データから 2 つのパラメータ α, β をどのように推定できるかについて説明する. 事前分布を設定するためにデータを使用することは不適切と思われるかもしれませんが, この例では, α, β などの分布パラメータを扱う階層的モデリング (5 章で紹介した) の好ましい手法の近似として推定する未知数としてこれを見ています.

このモデルでは, 任意の郡 j について観測されたカウント y_j は, この場合の Neg bin$(\alpha, \beta 10n_j)$ である予測分布 $p(y_j) = \int p(y_j|\theta_j)d\theta_j$. 付録 A から、この分布の平均と分散を見つけることができます.

$$\mathrm{E}(y_j) = 10n_j\frac{\alpha}{\beta},\ \mathrm{var}(y_j) = 10n_j\frac{\alpha}{\beta} + (10n_j)^2\frac{\alpha}{\beta^2} \tag{2.17}$$

これらは, 平均と分散の公式 (1.8), (1.9) を使うことで直接導くこともできる. 演習 2.6 を参照.

観測された平均と分散ををその期待値に合わせ α, β について解くことで, 事前分布のパラメータを導く. 年齢調整に対処しなければならず, 実際の計算はより複雑であり, 割合 $\frac{y_j}{10n_j}$ の平均と分散を扱う方が効率的です.

$$\mathrm{E}\left(\frac{y_i}{10n_j}\right) = \frac{\alpha}{\beta}, \ \mathrm{var}\left(\frac{y_j}{10n_j}\right) = \frac{1}{10n_j}\frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} \tag{2.18}$$

年齢調整に対処した後, 最後の式の $\mathrm{E}\left(\frac{1}{10n_j}\right)$ の代わりに値 $\frac{1}{10n_j}$ のサンプル平均を使い, 観測されたモーメントと理論的なモーメントを等しくし, $\frac{y_j}{10n_j}$ の値の平均を $\frac{\alpha}{\beta}$ に, $\frac{y_j}{10n_j}$ の値の分散を $\mathrm{E}\left(\frac{1}{10n_j}\frac{\alpha}{\beta} + \frac{\alpha}{\beta^2}\right)$ に設定する.

図 2.10 は, 元の癌率 θ_j に対する推定, Gamma(20, 430, 000) 事前分布とともに, 未処理癌率の経験的分布を示す. 生の割合の分布ははるかに広く, ポアソン変動と郡間の変動を含むため意味があります.

私たちの事前分布はこの例では合理的ですが, これを構成するこの方法は, 瞬間を合わせることによっていくらかぎこちなく, 一般的には適用するのが難しい場合があります. 5 章では, 階層モデルを使用して, より直接的なベイズ手法で, これおよび他の事前分布を推定する方法について説明する.

このモデルを改善するより重要な方法は, 郡レベルでがん率の変動を予測できる情報を含めることです. これは, 16 章で議論されているような, 階層的なポアソン回帰に向かってモデルを移動させるだろう.

2.8 無情報事前分布

事前分布に母集団の基礎がない場合は, 構築が困難であり, 事後分布において最小限の役割を果たすことが保証され得る事前分布が望まれてきた. そのような分布は, 参照事前分布 (reference prior distribution) と呼ばれることがあり, 事前密度は, 曖昧で, 平坦で, 拡散しているか, または無情報であると記述される. 無情報事前分布を使用する理論的根拠は, 多くの場合, 推論は, 現在のデータへの外部からの情報に影響されないように, 「データが, 自分自身について話すようにする」であると言われています.

関連するアイデアは, 事後分布を正規化 (regularize) するのに十分な情報を含む弱情報事前分布です. すなわち, おおよそ合理的な範囲内に保つことができますが, 潜在的なパラメータに関する科学的知識を完全に捉えることはありません.

適切な事前分布と不適切な事前分布

θ の $\mathrm{N}(\mu_0, \tau_0^2)$ 事前分布で, 既知の分散 σ^2 の正規分布の平均 θ を推定する問題に戻る. もし, 事前精度 $1/\tau_0^2$ がデータ精度 n/σ^2 に対して相対的に小さいならば, $\tau_0^2 = \infty$ としたとき事後分布は近似

$$p(\theta|y) \simeq \mathrm{N}(\theta|\bar{y}, \sigma^2/n)$$

である. これを別の言い方をすると, 事後分布は, $p(\theta)$ が $\theta \in (-\infty, \infty)$ の定数に比例すると仮定することに起因するものである. このような分布は厳密には可能ではない. 仮定される $p(\theta)$ の積分は無限大であり, 確率が 1 になるという仮定に反しているからである. 一般に, データに依存せず 1 に積分するならば, 事前密度 $p(\theta)$ を適切 (proper) と呼ぶ. ($p(\theta)$ が任意の正の有限値に積分されるならば, それは非正規化密度 (unnormalized density) と呼ばれ, 正規化され, 定数で乗算されて 1 に積分される.) この例では, 事前分布は不適切であるが, 少なくとも 1 つのデータ点が与えられれば事後分布は適切である.

無情報事前分布の 2 つ目の例として, スケーリングされた共役逆-χ^2 事前分布に関する, 既知の平均, 未知の分散を持つ正規モデルを考える. もし事前自由度 ν_0 がデータ自由度 n に比べ小さいならば, $\nu_0 = 0$ とすると事後分布は近似的に

$$p(\sigma^2|y) \simeq \text{Inv-}\chi^2(\sigma^2|n, v)$$

である. この事後分布の極限の式は, σ^2 に対し事前密度を $p(\sigma^2) \propto 1/\sigma^2$ として定義することで導くこともできる. ただこれは, $(0, \infty)$ で積分が無限になってしまい不適切である.

不適切な事前分布が適切な事後分布を導くことがある

上記の 2 つの例のいずれにおいても, 事前密度は, 適切な結合確率モデル $p(y,\theta)$ を定義する尤度と組み合わされる. しかしながら, ベイズ推論の計算を進め,

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

によって適切でない事後密度関数を定義できる. 上の例では (常にではない), 事後分布は実は適切である. すなわち, $\int p(\theta|y)d\theta$ はすべての y に対して有限である. 不適切な事前密度から得られた事後分布はかなり注意して解釈されなければならない. 事後分布が有限の積分と賢明な形を持つことを常に確認しなければならない. それらの最も理に適う解釈は尤度が事前分布を圧倒するような状況での近似である. ベイズ分析のこの点については 4 章でより完璧に議論する.

Jeffreys の不変原理 (Jefrey's invariance principle)

無情報事前分布を定義するために時々使用される 1 つのアプローチが, パラメータの 1 対 1 変換, $\phi = h(\theta)$ を考慮して, Jeffreys によって導入されました. 変数の変換によって, 前の密度 $p(\theta)$ は, 同じ信念を表すという観点から, ϕ に関する次の事前密度に等しい.

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta)|h'(\theta)|^{-1} \tag{2.19}$$

Jeffreys の一般原則は, 事前密度 $p(\theta)$ を決定するためのルールは, 変換されたパラメータに適用される場合、同等の結果をもたらす. すなわち, $p(\theta)$ を決定して (2.19) を適用することによって計算された $p(\phi)$ は, 変換されたモデル $p(y, \phi) = p(\phi)p(y|\phi)$ を直接使用して $p(\phi)$ を決定することで得られる分布に一致する.

Jeffreys の原理は, 無情報事前密度を $p(\theta) \propto [J(\theta)]^{1/2}$ として定義する. ここで, $J(\theta)$ は θ のフィッシャー情報である.

$$J(\theta) = \mathrm{E}\left(\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2 \middle| \theta\right) = \mathrm{E}\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} \middle| \theta\right) \tag{2.20}$$

Jeffrey の事前モデルがパラメータ化に対し不変であることを見るために, $\theta = h^{-1}(\phi)$ における $J(\phi)$ を評価する.

$$J(\phi) = -\mathrm{E}\left(\frac{d^2}{\log p(y|\phi)} d\phi^2\right) = -\mathrm{E}\left(\frac{d^2 \log p(y|\theta = h^{-1}(\phi))}{d\theta^2} \left|\frac{d\theta}{d\phi}\right|^2\right) = J(\theta) \left|\frac{d\theta}{d\phi}\right|^2$$

これより, 必要に応じて $J(\phi)^{1/2} = J(\theta)^{1/2} \left|\frac{d\theta}{d\phi}\right|$ となる.

Jeffreys の原則は多変量モデルに拡張することができますが, その結果はより議論の余地があります. ベクトルパラメータ θ の成分に対する独立した無情報事前分布を仮定することに基づくより簡単なアプローチは, Jeffrey の原理で得られるものとは異なる結果をもたらす可能性がある. 問題のパラメータの数が多い場合, 5 章で議論するように, 階層的モデルを優先して純粋な無情報事前分布を放棄することは有用であることがわかります.

二項パラメータに対する様々な無情報事前分布

対数尤度が

$$\log p(y|\theta) = \text{constant} + y \log \theta + (n-y) \log(1-\theta)$$

となる二項分布 $y \sim \text{Bin}(n, \theta)$ を考える 2 次微分のルーチン評価と $\text{E}(y|\theta) = n\theta$ の代入は, フィッシャー情報をもたらします.

$$J(\theta) = -\text{E}\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} \middle| \theta \right) = \frac{n}{\theta(1-\theta)}$$

Jeffrey の事前密度は $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ で, $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ 密度である. これと比較すると, $\theta \sim \text{Beta}(1, 1)$ で表せる, ベイズ-ラプラス一様事前密度を思い出す. 一方, 分布の指数関数表現の自然パラメータで一様事前密度は $p(\text{logit}(\theta)) \propto \text{constant}$ である (練習 2.7 参照). これは θ に対する不適切な $\text{Beta}(0, 0)$ に対応する. 実際には, $\theta \sim \text{Beta}(0, 0)$ から $\theta \sim \text{Beta}(1, 1)$ を得るためには, 1 回の成功と 1 回の失敗を前提としたものから後のものへのパスと同じであるため, これらの選択肢の違いはしばしば小さい通常, 観測の合計数の 2 分の 1 です. しかし, 不適切な $\text{Beta}(0, 0)$ 事前分布に注意する必要があります. $y = 0$ または n の場合, 結果の事後分布は不適切である.

ピボット量

二項分布やほかの単一パラメータモデルに対して, 異なる原理が (わずかに) 異なる無情報事前分布を与える. しかし 2 個のクラス, 位置あらメータとスケールパラメータに対して, すべての原理が次を満たす.

(1) y の密度が $p(y-\theta|\theta)$ が θ, y の自由な関数, 例えば $f(u)$, ここで $u = y-\theta$ であるのならば, $y-\theta$ はピボット量 (Pivotal quantity) で, θ は純粋な位置パラメータ (pure location parameter) と呼ばれる. そのような場合, θ に対する無情報分布が事後分布 $p(y-\theta|y)$ に対して $f(y-\theta)$ を与えるということが合理的である. すなわち, 事後分布のもとで, $y-\theta$ は, 分布が θ, y の自由となるようなピボット量であるべきである. この条件の下で, ベイズ則 $p(y-\theta|y) \propto (\theta)p(y-\theta|\theta)$ を使うことで, 無情報事前密度が θ 上で一様であることを意味する. すなわち, $(-\infty, \infty)$ で $p(\theta) \propto$ constant となる.

(2) y の密度が $p(\frac{y}{\theta}|\theta)$ が θ, y の自由な関数, 例えば $g(u)$, ここで $u = \frac{y}{\theta}$ であるのならば, $\frac{y}{\theta}$ はピボット量で, θ は純粋な位置パラメータと呼ばれる. そのような場合, θ の無情報事前分布が事後分布 $p(\frac{y}{\theta}|y)$ として $g(\frac{y}{\theta})$ を与える. 変数の変換により, θ が与えられた y の条件付き分布は, θ が与えられた u の分布の形

$$p(y|\theta) = \frac{1}{\theta}p(u|\theta)$$

と同じであるが

$$p(\theta|y) = \frac{y}{\theta^2}p(u|y)$$

で表現できる. $p(u|\theta), p(u|y)$ が $g(u)$ に等しいとすると, 等式 $p(\theta|y) = \frac{y}{\theta}p(y|\theta)$ を持つ. そして, この場合, 参照事前分布は $p(\theta) \propto \frac{1}{\theta}$ もしくは同じであるが, $p(\log\theta) \propto 1$ もしくは $p(\theta^2) \propto \frac{1}{\theta^2}$ である.

ピボットのサンプリング分布を事後分布として使用するこの手法は, 階層的なノーマルモデルなどのより複雑な例で十分統計量に適用できる.

これらの原則さえ, 不適切な事後分布につながる可能性がある事前分布を示唆する重要な意味で, いくつかの問題で誤解を招く可能性があります. 例えば, 5.4 節で論じるように, 一様事前密度は階層分散パラメータの対数に対しては機能しません.

無情報事前分布に関する難しさ

無情報事前分布の選択は次を含むいくつかの問題がある.

1. 常にあいまいな事前分布を検索するのは間違っているようです. 尤度が与えられた問題において本当に支配的であれば, 比較的平坦な過去の密度の範囲の中から選択することは重要ではない. 参照事前分布として特定の仕様を確立することは, 自動的かつおそらく不適切な使用を促すようです.

2. 多くの問題では, 漠然とした事前分布のための明確な選択肢はありません. これは, 1 つのパラメータ化で平坦または一様な密度が別のものにはないためです. これはラプラスの原理が不十分であるという原則, すなわち原理が適用されるべきスケールについての不可欠な難しさです. 例えば, 上の正規分布の平均 θ における合理的な事前密度は一様であり, σ^2 については密度 $p(\sigma^2) \propto 1/\sigma^2$ は妥当と思われる.

しかしながら, もし $\phi = \log \sigma^2$ と定義するならば, ϕ の事前密度が

$$p(\phi) = p(\sigma^2) \left| \frac{d\sigma^2}{d\phi} \right| \propto \frac{1}{\sigma^2} \sigma^2 = 1$$

すなわち, $\phi = \log \sigma^2$ 上で一様である. 離散分布では, 結果を等確率の原子に分割する方法を決定するという類似の難しさがあります.

3. 7.3 節で議論するように, 不適切な事前分布を持つ一連の競合モデルを平均化すると, さらに困難が生じます.

それにもかかわらず, 無情報かつ参照事前密度は, 事後密度が適切であることを確認するための数学的作業を実行したい限り, 確率分布としての実際の事前知識を定量化する努力をする価値がないように思われる場合に, 利便性のモデリングの前提条件に対する事後的な推測の感度を決定することである.

まとめ 2.8

・参照事前分布

…事前分布に母集団の基礎がない場合は, 構築が困難であり, 事後分布において最小限の役割を果たすことが保証され得る事前分布が望まれてきた. 「データが, 自分自身について話すようにする」

・弱情報事前分布

…事後分布を正規化するのに十分な情報を含むもの.

・適切, 不適切

…積分して 1 になる, または発散するような分布関数のこと.

・非正規化密度

…積分値が正の有限値になる分布関数のこと.

・正規モデル (ただし, 平均既知, 分散未知)

…無情報事前分布を, 逆-χ^2 分布とすると, ν_0 とすると, 事後分布は

$$p(\sigma^2|y) \approx \text{Inv-}\chi^2(\sigma^2|n, v)$$

この式は, 事前密度を $p(\sigma^2) \propto 1/\sigma^2$ とすることでも導けるが, これは不適切な事前分布である.

・Jeffreys の不変原理

…無情報事前密度を $p(\theta) \propto [J(\theta)]^{1/2}$ として定義する. ここで, $J(\theta)$ は θ のフィッシャー情報である.

$$J(\theta) = \mathrm{E}\left(\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2 \middle| \theta\right) = \mathrm{E}\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} \middle| \theta\right)$$

・二項パラメータに対する無情報事前分布

…Jeffreys 事前分布は $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ で, Beta$(\frac{1}{2}, \frac{1}{2})$ 密度である.

ベイズ-ラプラス一様事前密度は $\theta \sim$ Beta$(1,1)$ で表せる.

指数関数表現の自然パラメータで一様事前密度は $p(\mathrm{logit}(\theta)) \propto$ constant である. 不適切な Beta$(0,0)$ に対応する.

・ピボット量

…y について直接考えるのではなく $y-\theta$ を考えたりする. この量のこと.

・無情報事前分布に関する難しさ

1. 常にあいまいな事前分布を探すのは間違っている.
2. 多くの問題では, 事前分布についての明確な選択基準がない.
3. 不適切な事前分布を持つ一連の競合モデルを平均化するのは難しい.

2.9 弱情報事前分布

事前分布を適切であるがそれが与える情報が実際の事前知識が利用可能な時より意図的に弱いように設定されているならば弱情報的 (weakly informative) のように特徴づけられる. 具体的な例の文脈の中でこれについてさらに議論するが, 一般的に問題は弱情報的なモデルを可能にする自然な制約があります. 例えば, 対数またはロジスティックスケールの回帰モデルでは, バイナリまたは標準偏差が 1 になるようにスケーリングされた予測を使用して, 影響の大きさが 10 未満であるほとんどの応用に対して, 対数スケールで 10 の違いは期待値を $\exp(10) = 20,000$ の因子で変化させ, logit スケールでは $\text{logit}^{-1}(-5) = 0.01$ の確率を $\text{logit}^{-1}(5) = 0.99$ にシフトする.

この無知をモデル化しようとするのではなくて, 私たちはほとんどの問題において, 事後分布が意味を成立させるのに十分な少量の現実世界の情報を含む弱情報的事前分布を使用することを好む. 例えば, 2.1, 2.4 節からの性比の例で, 例えば, $\text{N}(0.5, 0.12)$, または共役分布, $\text{beta}(20, 20)$ の数学的利便性を保つために, 0.4 から 0.6 の間に集中した事前分布を用いることができる. 2.5 節からの正規平均の推定の一般的な問題 $\text{N}(0, A^2)$ 事前分布は弱情報的であり, A は問題の文脈に依存する大きな値に設定される.

ほとんどすべての実際の問題では, データアナリストは, 統計モデルに便利に含めることができるよりも多くの情報を持っています. これは, 事前分布だけでなく尤度の問題でもあります. 実際には、多くの理由から常に妥協があります. モデルをより便利に記述すること. 確率的な形で正確に知識を表現するのは難しいかもしれないからです. 計算を単純化する. 信頼できない可能性のある情報源を使用することを避けることができます. 最後の理由を除いて, これらはすべて便宜のための議論であり, 私たちがより正確であれば, 答えがあまり変わっていないという主張によって最も正当化されています. 無情報事前分布の選択が差異を生むデータが少ない場合, 5 章で議論するように, おそらく階層モデルを使用して関連情報を事前分布に入れるべきである. 後の章の例では, 尤度や事前分布の精度と利便性の問題に戻ります.

弱情報事前分布の構築

事実上すべての統計モデルが弱情報的であると主張するかもしれない. モデルは入力の選択と機能の組み合わせでのみ情報を伝えるが, 被験者に関する以前のすべての信念を確率分布のセットにエンコードすることは可能ではなく, おそらく望ましくない. そのことを念頭に置いて, われわれは弱い情報提供者を設定するために, 2 つの異なる方向から問題に進む 2 つの原則を提示します.

・無情報事前分布のバージョンから始め, 推論が合理的になるように制約されるように十分の情報を付け加える.

・強い情報事前分布から始め, 以前の信念における不確定性と, 新しいデータへの歴史的な事前分布の適用可能性を説明するためにそれを広げる.

これらのアプローチのどちらも純粋ではありません. 最初のケースでは, 出発点として使用された, 未知の事前分布が実際には強すぎるということが起こり得る. 例えば, U(0, 1) の事前分布が何らかのまれな疾患の確率に割り当てられている場合, 弱いデータの存在下では, 確率は総じて過大に評価される可能性がある ($n = 100$ のケースのうち $y =$ 真の罹患率は 10,000 人中 1 人未満であることが知られている), この事件の後部がその低い範囲に集中するような適切な弱い情報が前もってある.

第 2 のケースでは, 強く有益であると考えられる事前分布は, 実際にはある方向に沿って弱すぎるかもしれない. これは, 事後推論があいまいである場合には, 事前犯罪者をより正確にする必要があると言っているわけではありません. 多くの場合, 私たちの最善の戦略は, 単に私たちが持っている後天性の不確実性を認めることです. しかし, 私たちは, 実質的な事前知識が利用可能であるときに, デフォルトの無情報モデルによって制約を受けてはいけません.

しかし, たとえ明らかに事後推論を改善できたとしても, 関連情報を使用しないことを推奨できる場合は, 設定があります. ここでの関心事は, 公平の観点からしばしば表現され, 対称原理として数学的に符号化されているので, 事前分布は所定の方向で推論を引くべきではない. 例えば, 彼女がかなり確信しているという効果を研究している実験者が肯定的であると考える. おそらく, 彼女の事前分布は, 適切な規模で $\mathrm{N}(0.5, 0.5)$ であろう. このような前提は, 現在の科学的情報を考慮すると明らかに合理的かもしれないが, 科学者の理論をテストするために設計された実験の分析の一部であれば, 潜在的に危険に見える. 何かがあれば, より高い基準の証明を要求するために, 実験者の仮説に傾いている事前分布が欲しいかもしれない.

結局のところ, そのような懸念は, 決定分析や科学プロセス全体のモデルの一部に包含でき, また, されるべきであり, 影響の大きさを過大評価し, そのパターンに過度に反応する損失に対する大規模かつ実質的な影響を早期に特定する利益チャンスに起因する可能性があります. しかしながら, その間, 統計的推論が影響の証拠として, 未来の意思決定の手引きとして扱われ, この目的のためには, 単一取り扱いの効果の事前分布のためにモデルが特定の制約, 例えば 0 についての対称性を有することをモデルに要求することは意味をなさない.

まとめ 2.9

・弱情報事前分布

…適切であるがそれが与える情報が実際の事前知識が利用可能な時より意図的に弱いように設定されている事前分布のこと.

・弱情報事前分布の構築

1. 無情報事前分布から始め, 推論が合理的になるように制約されるように十分な情報を付け加える.

2. 強い情報事前分布から始め, 事前の信念における不確定性と, 新しいデータへの事前分布の適用可能性を説明するためにそれを広げていく.

ただ, 両方ともどちらのアプローチも純粋ではない.