Chapter 6. Prior

2010/5/29 (Sat.) 飯島勇人*†

この章の目的

1. 事前分布が事後分布に与える影響
2. 種々の無情報事前分布の性質と問題点

訳語（なんとなく普通でない訳しかたをしているところをあげておきます）

- prior: 事前分布
- posterior: 事後分布
- improper prior: 非正則事前分布
- objective: 客観
- subjective: 主観
- "": 「」で表記

ベイズ推論の哲学上の魅力-あらゆる不確実性を表現するのに確率分布を用いるという固有の性質、単純さと正確さ-は、未知のパラメータに事前分布を設定する必要性から、一部では無価値なものとされている。事前分布の選択に関連した主観性という（ベイズ統計の）雰囲気は、今日科学者によるベイズ推論の利用の拡大を妨げる最も大きな障壁である。

事態を悪化させているのは、多くのベイズ論者が、定義の基本となるキログラム基準のような客観的な確率のようなものは存在しないと主張していることである。全ての問題は一致性である: 他者には他者の確率があり、私には私の確率があり、問題なのは我々がそれらを操作するという規則に同意していることである。そのような論者は、確率に関連して理学的または実質的な真実が存在するのか、あるいはその必要性があるのかと問いかけ、主観確率を擁護する[1]。そのため「主観性」という言葉は、不誠実という非科学的な陰という要素は持たず、技術的な意味でベイズ解析と結びつけられてきた。我々は事前分布を選ばなければならないのであり、技法的な意味でなく主観性を汚さないようにするためにどのように進めるべきなのだろうか？議論を、以下の観察から始めよう。

$$\text{Inferential Basis} = \text{Data} + \text{Prior Knowledge} \tag{6.1}$$

この記憶法は、ベイズやそれ以外の方法による全ての推論を表現している。「データ」が単なる数字またはシンボルのリスト以上のものである状況であるなら、事前の知識は必然的に用いられる。

* 山梨県森林総合研究所森林保護科研究員（注！（独）森林総合研究所とは一切関係ありません）
† 連絡先: hayato.iijima@gmail.com または http://www7.atwiki.jp/hayatoiijima/

[1] そのような哲学的な熟考をしてみたい読者には、I.I.Good の興味深い評論集（Good, 1983）を薦める。

適切な推論では常にこの仮定が認められ熟考され、自己を欺こうとすることや偏った主張を抑制している。ベイズの枠組みでは、パラメータの事前分布が仮定されている; 分析者は選択した事前分布を報告することで、うそを装っていてもそれを回避することができる。

そのため、ベイズ推論はデータ解析ではほぼなく、データと事前分布の分析であることを堂々と認識し認めるべきである。事前情報が推論に与える影響は、単に複数の事前分布による分析を実行することで簡単に十分評価できる。

ある特定の値の付近に限定しない事前分布を与えた場合、データ数が増加するとデータが事前分布を覆い尽くす傾向がある。もし過去のデータを持っているのであれば、事前分布の選択は問題とならない。我々は、科学者がデータから言えることが制限されている（限られたデータってこと？）状況での洞察に関心があると言うことを知っている。もしたくさんのデータがあれば複雑なモデルを適用しようとし、そのような状況でも事前情報は推論にある程度の影響を与える。一方で、フリーランチのようなもの（統計的手法）は存在しない: 全ての統計的推論の手法は、データをほとんど持っていないときには欠点がある。

事前分布が推論に影響を与えるので、時にはベイズ分析は正確で望ましいものである。過去の研究あるいは共通認識かによらず、パラメータの事前情報は確率分布によって定量化されるだろう。この（確率）分布は informative prior（主観事前分布）と表現される。事後推論はデータの情報に事前情報を適用した際の決められた手続きである。例を 6.1 で示す。

しかしながら実際にはしばしば主観事前分布は用いられず、分析者はデータそのものに語らせようとし、無情報事前分布が用いられる。彼らは可能な限り式 6.1 の右辺の影響を最小化しようとし、「客観的なベイズ分析」を行おうとする。

おそらく驚くことだろうが、（データの形があれば？）自動的に決定され、単純で万人に認められた方法は存在しない。事前の情報がないことを表現する事前分布を定義することは、思っているよりも困難である。我々はなぜ、そして客観的なベイズ分析を行う一般的な原則を 6.2 で説明する。

6.1 An example where prior matter

ハマヒメドリ（*Ammodramus maritinus nigescens* の亜種ですが以下では単にハマヒメドリと表記します）は Florida 州の南部中央の海岸沿いの塩湿地で生育していたが、Kennedy Space Center において実施された蚊の個体群を制御するための塩湿地管理事業によって絶滅した。1987 年の絶滅につながったこの悲しい出来事は Walters（1992）によって語られている。

1979 年、飼育下で増殖させようとし、残っていた 6 羽を捕獲しようとする試みが始まった。5 羽が捕獲されたが、全てオスだった。そこでおそらく我々が問おうとするのは、残りの逃げた 1 羽がメスである確率はどれぐらいなのだろうか？

サンプルサイズが少ない状況では、事前分布の選択がモデルのパラメータに相当敏感に影響すると予測するだろう。事後推論を行うためには、事前情報の仮定とその論理的な因果関係を説明しなければならない。我々は持っているデータのモデルとして用いるであろう超幾何分布によるベイズ推論から始める。

6.1.1 Hypergeometric distribution

N 個の玉が入っているとわかっている箱から、玉を n 回非復元抽出で取り出すことを考える。n 個玉を取り出した内 m 個が赤玉であった; 我々は箱の中にある全赤玉の数 M を知りたい。m の確率分布は超幾何分布であり、その確率密度関数は、

$$f(m|M,N,n) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \tag{6.2}$$

である。推論は事後分布に基づくであろう。

$$\pi(M|m,N,n) \propto f(m|M,N,n)g(M|N,n) \tag{6.3}$$

事前分布 $g(M|N,n)$ が必要なことに注目すべきである; これはおそらく n には依存せず N に依存すると考えられるので、この事前分布は $g(M|N)$ とする。

M に関する事前情報が全くない状況では、おそらく離散一様分布（$M = 0, 1, 2, ..., N$ の時 $g(M|N) = 1/(N+1)$）を選択するだろう。そのため、事後分布は、

$$\begin{aligned}\pi(M|m,N,n) &= \frac{\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}\frac{1}{N+1}}{\sum_{H=0}^{N}\frac{\binom{H}{m}\binom{N-H}{n-m}}{\binom{N}{n}}\frac{1}{N+1}} \\ &= \frac{\binom{M}{m}\binom{N-M}{n-m}}{\sum_{H=m}^{N-(n-m)}\binom{H}{m}\binom{N-H}{n-m}}\end{aligned} \tag{6.4}$$

$\binom{N}{n}$ と $\frac{1}{N+1}$ がなくなることに注目せよ; これらは M の確率分布として考えると定数である。また、$H < m$ なら $\binom{H}{m}$ であるから、$H = m$ から加法を開始することにも注目すべきである（分母）。これは $m > M$ の場合に $f(m|M,N,n) = 0$ となることを反映している: 我々は最初にあった以上の赤玉を取り出すことができない。同じように、加法は $N-(n-m)$ で終了する; 我々は少なくとも赤でない玉が $n-m$ 個あることがわかっている。

試しの計算結果が Table 6.1 に示されている。箱には $N = 30$ 個の玉があり、我々は $n = 20$ 個の球を取り出し、$m = 18$ 個が赤玉であった。箱から玉を取り出す前、我々は赤玉の数に関して全く知識がなく、$N = 0, 1, 2, ..., 30$ に等確率を与えた（一様分布）。20 個の球を取り出し 18 個が赤玉であったら、我々は赤玉が 18 個以下であることはなく 28 個以上ではないことを確信する。事後分布を用いると、我々は箱の中に少なくとも 5-10 個の赤玉がある確率が 99.4% であると結論する。なぜなら、Pr(#Red remainig ≥ 5) = Pr(M - m ≥ 5) = Pr(M ≥ 23) = 1 - 0.6% となるからだ。

6.1.2 Dusky seaside sparrow

最後の 6 羽目のハマヒメドリの捕獲確率が（他の 5 羽と？）等しいと仮定すると、式 6.2 で与えられる確率密度関数に $n = 5$ と $N = 6$ を与えることで、超幾何分布の変数 m として捕獲された雄

鳥の数をモデル化することができる。雄鳥の数 M については離散一様分布を用いると、式 6.4 から $M = 5$ の時は $\pi(M|m = 5, N = 6, n = 5) = 1/7$、$M = 6$ の時は 6/7 となる。選択した事前分布の元では、残る 1 羽が雄である可能性のオッズ比は 6 : 1 である。

この場合、M に関する別の事前分布も妥当である: Fisher（1935）はノンパラメトリックなモデルについて「無知による間違った仮定は無害ではない; それはしばしば明らかに不合理なものとなる」と表現している。例えば、二項分布を用いた事前分布 $g(M|N) = \binom{N}{M}(1/2)^M(1-1/2)^{N-M} = \binom{N}{M}(1/2)^N$ は性比が 1:1 で生じると言う予測のより現実的な描写かも知れない。この事前情報を式 6.3 に用いると、事後分布[2]は、

$$\pi(M|m, N, n) \propto \binom{M}{N}\binom{N-M}{n-m}\binom{N}{M} \tag{6.5}$$

$m = n = 5$ および $N = 6$ とすると、式 6.5 の右辺は $M = 5$ と $M = 6$ の両方の場合 6 になる。Since there are the only possible values、$M = 5$ と $M = 6$ の事後確率は同じ、すなわち 50% となる。捕獲できなかった 1 羽が雌である可能性は、事前分布の選択により 1/7 から 1/2 に増加する。

この 2 つの事前確率の違いはいくつかの観点から意味がある。M に関するそれぞれの事前分布は、性別を N 羽の鳥について生起確率 $p = \Pr(\text{Male})$ である独立したベルヌーイ試行と扱った結果であると考えることができる。もし p について知識がなく未知のものとして扱うなら、p は 0 から 1 を等確率で取る一様分布としてモデル化され、その結果 M は $M = 0, 1, 2, ..., N$ について以下のような分布となる。

$$\begin{aligned} g(M|N) &= \int_0^1 \binom{N}{M} p^M (1-p)^{N-M} dp \\ &= \frac{\binom{N}{M}\Gamma(M+1)\Gamma(N-M+1)}{\Gamma(N+2)} = \frac{1}{N+1} \end{aligned}$$

これは最初の分析で用いた事前分布である[3]。二番目の分析では、全ての鳥について $p = 0.50$; それぞれ雄である確率が 0.5 としている。p に関する事前分布の黙示された違いは捕獲されていない鳥が雌である事後確率の違いを説明する。Y は逃げている鳥が雌である事象を表す変数であるとする。すると、

$$\Pr(Y = 1) = \mathrm{E}(\Pr(Y = 1|p)) = \mathrm{E}(1-p) = 1 - \mathrm{E}(p)$$

捕獲された 5 羽のハマヒメドリの性別は、パラメータ p について事前分布 $Be(1, 1)$ から事後分布 $Be(1 + 5, 1 + 0)$ に更新される。この最初の分析方法では、$Be(6, 1)$ の平均は $\mathrm{E}(p)$=6/7 であるので、$Y = 1$ である事後確率は 1/7 である。2 番目の分析では、p は捕獲されるされないにかかわらず全てのハマヒメドリで等しい; 5 羽の捕獲されたハマヒメドリは捕まっていない 1 羽について全く情報を与えない。そのため 2 番目の分析では、$Y = 1$ である事後確率は 1/2 である。

---

[2] $\binom{N}{M}$ と $(1/2)^N$ は M という変数に関しては定数であるので比例定数に吸収される

[3] この積分は「統計学者のように積分する」ことで容易に計算される: 被積分関数は既知の密度関数、この場合はベータ密度関数とほぼ同一であると認められ、積分の結果 1 となるような適切な定数を乗ずる

しかし、どちらの答え、Pr(Female) = 1/2or1/7 が正しいのだろうか？答えは「どちらでもなく」「どちらも正しい」である。どちらでもない、とは、逃げている 1 羽に関する追加の知識があれば事前情報が異なり、異なる結論を得るという意味である。どちらも正しい、とは、それぞれの分析が数学的に正しく、きちんと計算されて得られたという意味である。それぞれの値が得られるのに際しおかれた仮定について難癖をつける者がいるかも知れないが、（事前分布の？）前提条件については明確に数学的な評価が行われているので、結果について異論を挟む余地はない。今回の事例に関する我々の立場は、1/7 がより現実的であると思われるということである。というのは、非常に強いストレスの影響下にある個体群で性比が一定であると仮定する理由がないからだ。

全ての個体の捕獲確率が等しいという仮定の影響について、以下のようにして検証することができる: $[M|N,\phi] = B(N,\phi)$（前のように $\phi = 1/2$ あるいは $\phi \sim U(0,1)$ が既知とする）とし、$[m|M] = B(M,\pi_M)$ および $[f|F] = B(F,\pi_F)$（$F = N - M$ とする）とする。データは $m = 5$ および $f = 0$ である。以前の分析では $\pi_M = \pi_F$ と仮定しており、$[m|n = m + f, M]$ は超幾何分布であるという事実を利用している。その代わりに、π_F と π_M に独立な一様分布を当てはめる。以前の解析と新しい解析の結果は Table 6.2 に要約されている。

これらの推論の違いをどのように理解したらいいのだろうか？最初に、データは、捕まっていない 1 羽が雌である確率に対して、$\phi = \Pr(\text{Male}) \geq 5/6$ である点に注目すべきである。そのため、$\phi = 0.50$ とすることは捕まっていない 1 羽が雌である事後確率を増加させる。データは、$\pi_M = \Pr(\text{Captured}|\text{Male}) \geq 5/6$ ということも示唆している。$\pi_M = \pi_F$ とすることは、雌の捕獲確率を著しく高めることを意味している。そのため、$\pi_M = \pi_F$ という制限をなくすことは、雌の捕獲確率を減少させ、そのため残りの 1 羽が雌である確率を増加させる（？）。

Table 6.2 にあるような、比較できない 4 つの答えは悩ましいものであり、特にデータがモデルによる予測を行うには少なすぎると考えられる場合は特にそうである[4]。一方で、それぞれの計算結果はあるデータと仮定の元での正確な結果の要約である; no ambiguity has been built in through the use of dubious analytical approximations. これらの結果の違いはデータが制限されていることを際立たせ、様々なモデルの仮定について我々に考えさせる。何人かは手を挙げ、なんの結論も得ることができないというかも知れない。また別の者たちは、様々な結果に重みをつけ、結果を非公式に組み合わせようとするかも知れない: 「1/7 には 70%、7/19 には 15%、1/2 には 10%、7/9 には 5%、これらの合成値はおおよそ 1/4 という結論に到達する」。推論の重み付けのより正式な方法は 7 章で議論する。

あれやこれやとあるが、分析は、逃げているハマヒメドリの性別に関して引き出そうとする結論は、事前分布の信念（考え方？）に大きな影響を受けると言うことを明確に示している。ここで、無情報ベイズ推論に目を向け、得ようとする結論に対して事前分布の影響をどうやって最小限にするかについてみてみよう。

[4] Table 6.2 の要約された結果は、4 つのモデルだが事前分布の指定によって区別可能であると考えられる。$\phi = 1/2$ のモデルは退化した事前分布（ϕ に点の事前分布を与えている）であり、$\pi_M = \pi_F$ のモデルも同様である（ある単位平面において $y = x$ の線に集中する混合分布）

6.2 Objective baysiean inference

あるデータ X が与えられたときのパラメータ θ の事後分布は、

$$[\theta|X] \propto [X|\theta][\theta] \tag{6.6}$$

これは尤度と事前分布の積である。どのようにして事前分布の信念（考え）の効果を最大化しているのかは容易に見て取れる: 尤度を選び、θ の事前分布を非常に限定されている範囲で選択する。$\epsilon > 0$ ようなある値について、$[\theta_0 - \epsilon, \theta_0 + \epsilon]$ という範囲の一様分布を θ の事前分布としよう。式 6.6 の右辺は（上記の値の）範囲外では 0 となるから、事後分布 $[\theta|X]$ は（事前分布と）同じ範囲である。ϵ について非常に小さい値を選べば、データ X に関わらず、事後分布を θ_0 に非常に近い値に制限することができる。

事前の知識の一部分は疑いようのないものであり、それはデータが示唆することを無視することになると考えられるが、通常はデータから何かを知りたいのである。現実には、事前分布とデータの影響の不均衡を解消しようとするのであれば、より極端な望ましい方法として、事前の知識の影響を最小化してしまうのが典型的である。事前の知識が最終的な推論にほとんど影響しないような「客観的な」ベイズ解析はどのように行ったらよいのだろうか？無情報な事前情報はどのように定義すればいいのだろうか？

第一歩は、尤度 $[X|\theta]$ を構築するために可能な限り「事前の情報」を制限することに同意することである。次に、θ に非常に狭い一様分布を用いる代わりに、非常に範囲が広い一様分布を選ぶ。もしモデルのパラメータの空間が有限であるなら（例えば、ベルヌーイ試行における成功確率 p について $0 < p < 1$、角度測定における ψ について $0 < \psi < 2\pi$)、全域を覆う一様分布を指定することができる。これは θ について（事前情報について）無知である場合の指定方法して適切であるように思われる: あらゆるある一つの値も他の値より確からしいと言わないのである。Thomas Bayes（1973）は（ベイズ理論を発表した？）原著論文の中で、逆確率関数（？）について事前情報がないことを表現する方法として一様分布を用いており、そのため（一様分布を）選択することの適切さは「ベイズの仮定」として知られている。一様分布の事前情報はある定数 c について $[\theta] = c$ となる分布であり、式 6.6 から事後分布は尤度に比例することに注目せよ。特に、事後分布のモードは最尤推定値となる。

一様事前分布は客観的なベイズ分析にとって興味のある基準であるように思えるかも知れないが、小さな欠点を指摘しておかなくてはならない。第一に、パラメータ θ について一様事前分布を設定することは、変換したパラメータ $\psi = g(\theta)$ についてはほぼ常に一様分布ではないことを意味している。そのため、異なる者がパラメタリゼーションの方法が異なる同じモデルを同じデータに適用すれば、得られる答えは異なる。パラメータ θ について無知であるという表現は、パラメータ ψ については主観事前分布になっている。そのため、事前分布は「変換に対する不変性」はない。この問題は通常聞こえほど悪い問題ではない; 6.2.1 節でそれを示す。

一様分布を無情報事前分布として用いるもう一つの問題は、θ の範囲が無限大になってしまうか

も知れない（例えば平均が μ の正規分布の場合は $-\infty < \mu < \infty$）ことである。無限の範囲を持つ非 0 の定関数を積分すると 1 ではなく ∞ となってしまうので、無限の範囲を持つ一様分布は不可能である; $[\theta] \propto c$ となる事前情報は「不適切（非正則）」と言える。この非正則事前情報は、厳密には正しくなくないが、しばしば意味のある結果を得られる方法によって事後分布として扱うことができる[5]。

$$f_X(\theta) = \frac{[X|\theta]}{\int [X|\theta] d\theta}$$

$f_X(\theta)$ に基づいた結果は通常非常に大きいがしかし有限な範囲を持つ正則な事前分布によって得られた結果と一致する。非正則事前分布については 6.2.2 節で議論する。

上述のような方法によって無情報事前分布を構築し実行するために、多くの努力が（なされ）拡張されてきた（Kass and Wasserman, 1996）。Bernard（1979）は、事前分布から事後分布に対するカルバックライブラーダイバージェンスを最小化するような事前分布を *reference prior* と定義した。Bernard の reference prior は数学的に複雑である; しかし事後分布が漸近的に正規分布となることが保証されるという条件の下では、6.2.3 節で記述する Jeffreys 事前分布と一致する。一様分布とは異なり、Jeffreys 事前分布は変換不変性がある。

6.2.1 Uniform priors and transformation invariance

データモデルをある p について $0 < p < 1$ となる二項分布 $X|p \sim B(N, p)$ とする。一様分布 $[p] = U(0, 1) = Be(1, 1)$ がパラメータ p の無情報事前分布として適切な候補と思われる: p について取り得る全ての値が全て同じぐらりありそうだと仮定する。しかし、$\psi = p^2$ であるパラメータについて問われたとする。パラメータ ψ もまた 0 から 1 の間を取るが、$[p] = U(0, 1)$ であれば、

$$\Pr(\psi \leq t) = \Pr(p^2 \leq t) = \Pr(p \leq \sqrt{t}) = \sqrt{t}$$

そのため、ψ の確率密度関数は

$$f(t) = \frac{d\sqrt{t}}{dt} = \frac{1}{2\sqrt{t}}, 0 < t < 1$$

となり、Fig. 6.2 に示されている。ψ の分布は一様ではない; それは $Be(1/2, 1)$ である。$\psi \leq 0.25$ の確率は 50% あり、ψ の期待値は 1/3 である。そのため、「無情報」が「一様」と定義されると、p についての無情報事前分布は p^2 については主観事前分布となる。「p については何も知らないが、p^2 については何かを知っている」と言うことは人騒がせである。

上述で示したように、一様な事前分布は変換に不変ではない[6]。これは恐ろしく深刻な問題ではない。$\psi = p^2$ の事前分布として $U(0, 1)$ を選んだとする。上記の理由から、p の事前情

[5] この間数はデータ X とおそらく確率分布から定義されているが事後分布ではなく、それに関連した正則な事前情報は存在しないことを強調するため、$f(\theta|X)$ ではなく $f_X(\theta)$ という単語を用いた。

[6] 頻度論における不偏性の基準もまた変換に不変ではない。例えば σ^2 の推定値としての s^2 の不偏性は、σ の推定値としてはバイアスがある。この証明は以下のようである: $0 < \mathrm{Var}(s) = E(s^2) - E(s)^2 = \sigma^2 - E(s)^2$ であるから、$E(s)^2 < \sigma^2$ であり、$E(s) < \sigma$

報としては $Be(2,1)$ を導入していることが証明されている。この結果としての事後分布は、$Be(X+1, N-X+1)$ ではなく $[p|X] = Be(X+2, N-X+1)$ である。$X = 15$、$N = 30$ の場合にこれらの 2 つの事後分布をプロットしたのが Fig. 6.3. である; 一様分布の場合の 95% 最高事後密度区間は（0.33, 0.67）であり、$Be(2,1)$ の事前分布の場合は（0.35, 0.68）である。そのため、非常に少ないデータセットであっても、推論は事前分布の選択に影響するほど敏感ではない。

6.2.2 Improper priors

一様事前分布は、パラメータが有限の範囲を持つ時に、適切と考えられる値が特に存在しないときに有効である。この考えを無限に拡張することはよいように思われるが、無限の範囲を持つ一様分布は不可能である。*improper prior*（非正則事前分布）という考えを導入する必要がある。

データモデル $p(X|\theta)$ と事前分布 $\pi(\theta)$ が与えられると、事後分布は

$$f(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{\int p(X|\theta)\pi(\theta)d\theta} \tag{6.7}$$

となる。ここで、式 6.7 の $\pi(\theta)$ の代わりに非負の関数 $g(\theta)$ を代用するのは意味のあることである。$g(\theta)$ は無限の範囲の積分を持ち確率分布になれない（積分して 1 にならない？）が。関数 $g(\theta)$ は非正則事前情報と称される。たとえ $\int g(\theta)d\theta = \infty$ であろうとも $\int p(X|\theta)g(\theta)d\theta$ は有限となることは起こりうることであり、その結果は

$$f_X(\theta) = \frac{p(X|\theta)g(\theta)}{\int p(X|\theta)g(\theta)d\theta} \tag{6.8}$$

と定義される。これは完全に数学的に正しい確率分布で、まるで事後分布のように用いる。(improper prior の例として）無限の範囲の値の場合の一様分布をすでに示している。他の例は、共役事前分布を用いる状況で登場する。

Improper Priors in Conjugate Families: Binomial Success Rate

もしもある分布系の事前分布 $[\theta] = g(\theta|\psi)$ を尤度 $[X|\theta]$ と組み合わせ、事前分布と同じ分布系の事後分布が生成されるのなら、事前分布は尤度に共役であると言われる（4 章)。事前分布は超パラメータ ψ_0 を持ち、事後分布は超パラメータ ψ_1 を持つ。典型的には、ψ_1 は ψ_0 とデータ X から容易に計算できる。例えば、$p^X(1-p)^{N-X}$ に比例する二項尤度が与えられたときに、超パラメータ $\psi = (\alpha, \beta)$ を持つベータ分布属は共役である。更新する式は以下のようである。

$$\alpha_1 = \alpha_0 + X, \beta_1 = \beta_0 + (N - X) \tag{6.9}$$

章 4 において、共役性を計算効率の点から議論した。Conjugacy is also useful in that the updating formula may suggest a form for a noninformative prior. 例えば、ベータ分布のパラメータ α と β が式 6.9 に従って全体の成功と失敗数によって更新される。p の一様事前分布は $\alpha_0 = \beta_0 = 1$ によって得られる; これは 2 回の試行において 1 回成功するという事前の知識と等価であると解釈することができるかもしれない。なぜ知識がないと言うことをより強く示す基準を使

わず、0 回の試行で 0 回成功するということを示す事前分布 $\alpha_0 = \beta_0 = 0$ を設定しないのだろう。$Be(0,0)$ という事前分布は Haldane（1931）によって提唱されている。

唯一の問題は、$Be(0,0)$ のような分布が存在しないことである。超パラメータ α と β は正でなければならない。$Be(0,0)$ のようなものが存在したとするならそれは $g(p) = p^{-1}(1-p^{-1})$ に比例するものとなり、$[0,1]$ の区間で積分すれば 1 になるだろう。しかし、関数 $g(p)$ は $p \to 1$ または $p \to 0$ の場合に急速に無限大となり、以下の式のような結果となる。

$$\int_0^1 g(p)dp = \infty$$

結果的に、$g(p)$ を積分して 1 に調整するための定数 c がない; 無限の範囲を取る一様分のように、$g(p)$ は非正則な事前分布なのである。

しかしながら、式 6.8 のようにまるで正則な事前分布であるかのように $g(p)$ を扱うことができる。$X \neq 0$ または N と $f_X(\theta)$ が数学的に正しい分布関数であるとする; 非正則な事前分布は正則な事後分布として扱えるようなものとなる。この「事後分布」もまたベータ分布属であり、式 6.9 に従った超パラメータの更新がなされる; $\alpha_1 = 0 + X = X$ および $\beta_1 = 0 + N - X = N - X$。

この非正則な事前分布は、正則な事前分布が $Be(0,0) = \lim_{\epsilon \to 0} Be(\epsilon, \epsilon)$ という極限に向かったもの、結果として得られる事後分布は対応する事後分布が $Be(X, N-X) = \lim_{\epsilon \to 0} Be(X + \epsilon, N - X + \epsilon)$ という極限に向かったものであると考えるのが最適である[7]。詳細に多大な注意を向けないようにするため、$X = 0$ の場合を考えてみよ。$p > 0$ において事後分布は ϵ が小さくなると 0 に近づく。もし非正則な事前分布を用いて $X = 0$ を観測したとしたら、事後分布は $N = 1$ であろうと確実に $p = 0$ を示す。

以下では非正則な事前分布が非正則な事後分布を導くかも知れないことを示す。この問題は避けられないものではないが、非正則な事前分布は非常に有用かも知れない（？）。一つの例は、ランダムに得られるサンプルにおける通常の平均を推定する問題である。

Improper Priors in Conjugate Families: Normal Mean

$[\bar{X}|\mu] = N(\mu, \sigma^2/n)$ が与えられ、σ が既知であり、正規分布が μ の共役な分布であるとする。そのため $[\mu] = N(\mu_0, v_0)$ なら、事後分布は $[\mu|\bar{X}] = N(\mu_1, v_1)$ であり、

$$\mu_1 = (\frac{\sigma^2/n}{v_0 + \sigma^2/n})\mu_0 + (\frac{v_0}{v_0 + \sigma^2/n})\bar{X} \tag{6.10}$$

$$v_1 = \frac{v_0\sigma^2/n}{v_0 + \sigma^2/n} \tag{6.11}$$

[7] この非正則な事前分布に対する考えは、BUGS ソフトにおいて超パラメータ α と β がともに厳密に正でなければならないことを強いている。$\alpha = \beta = 0$ ではなく $\alpha = \beta = 0.001$ を選べばいいと思うかも知れない。しかしもし非正則な事前分布が非正則な事後分布を導くなら、非正則な事前分布は読者にこの本をほぼ確実に閉じさせないだろう。-結果として得られる事後分布はほぼ非正則であり、悪い推論にしかならない。

無限の分散を持つ非正則な事前分布は事前の情報がないと言うことの表現として非常に適切に思われる。この事前分布は正則な事前情報が $v_0 \to \infty$ という極限に向かったものである。式 6.10 および 6.11 を見ることで、非正則な事前分布 $N(\mu_0, \infty)$ が事後分布 $N(\bar{X}, \sigma^2/n)$ になったことがわかり、この事後分布から $(1-\alpha)100\%$ の信用区間が計算できる。

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

この信用区間は数字としては標準的な頻度論における信頼区間と同等である。2 つの区間が数字として一致することで、概念間の違いを不明瞭にするべきではない。ベイズ推論での解釈はあるパラメータ（が取り得る範囲）の確率であり、頻度論においては method の確率である。

6.2.3 Jeffreys priors

無情報の候補として一様分布を考えた。一様な事前分布が与えられると、事後分布は単に尤度に比例したものとなる; 事前情報がないことは、尤度以外に推論に何物ももたらさない。一様な事前分布はパラメータの範囲が有限の場合に意味があり、不適切な事前情報を用いることにはなるが無限の範囲を持つパラメータに自然に拡張される。

しかし、変換不変性の問題はどうだろうか？ 6.2.1 節の題材は、一様分布を用いて「無情報であること」を示せば、パラメータ θ の無情報事前分布は変換したパラメータ $g(\theta)$ にとっては主観事前分布となることを示している。そのため、同じモデルを（変換したパラメータで）再パラメタリゼーションすると、異なる推論につながるかもしれない。よいデータが与えられればその違いはわずかな物かも知れないが、それでもその問題は存在するのである; もしそのような気持ち悪さを回避できるような「無情報」の性質を定義できるのであれば、それはよいことであろう。

Jeffreys（1946）は賢い解決方法を提案し、彼の名前を持った事前分布の指定方法が誕生した。彼の解決方法を理解するために、変数変換の定理（2.2.4 節）と Fisher の情報量の定義と性質を思い出そう。

Fisher Information

$L(\theta|X) \propto [X|\theta]$ をある観察データ X の元での θ の尤度とし、スコア関数を θ について対数尤度 $\log L(\theta|X)$ について微分したものとする、すなわち、

$$S(\theta|X) = \frac{d \log L(\theta|X)}{d\theta} = \frac{dL(\theta|X)/d\theta}{L(\theta|X)} \tag{6.12}$$

Fisher 情報量 $I(\theta)$ は、θ が固定されていて一連の X が与えられたときのスコア関数の平方和の期待値として定義される[8]。それは、

$$I(\theta) = E_X(S(\theta|X)^2) \tag{6.13}$$

[8] Fisher（1925b）は Fisher 情報量を、パラメータについてデータの分布に関する二次導関数の期待値にマイナスをつけたもの、すなわち $-E(d^2 f(X|\theta)/d\theta^2)$ と定義した。これら 2 つの定義は緩やかな制限の元で一致する。

このモデルをパラメータ ψ で再パラメタリゼーションするとしよう。θ's から ψ's を計算できなければならないし、その逆もそうであるので、$\psi = \psi(\theta)$ および $\theta = \theta(\psi)$ と書くことができる。ψ についての対数尤度関数を計算するためには、θ についての対数尤度関数について $\theta = \theta(\psi)$ と代入するだけでよい。そのため、ψ についてのスコア関数は連鎖律によって以下のようになる。

$$S(\psi|X) = \frac{d \log L(\theta(\psi)|X)}{d\psi} = \frac{d \log L(\theta(\psi)|X)}{d\theta} \frac{d\theta}{d\psi} = S(\theta|X) \frac{d\theta}{d\psi}$$

式 6.13 を用い、$d\theta/d\psi$ がデータ X については定数であることを理解すると、Fisher 情報量は以下のようになる。

$$I(\psi) = I(\theta)(\frac{d\theta}{d\psi})^2 \tag{6.14}$$

なぜ $I(\theta)$ が「情報」と称されるのか？式 6.12 から、X が固定されていればスコア関数は θ が変化する際の尤度の変化率である。緩やかな数学的条件の下で、（データセット X の元での）変化率の期待値は 0 となるため、Fisher 情報量はスコア関数の分散である。そのようなもののため、Fisher 情報量は観察値が尤度の変化に与える影響を測るものである。

Jeffreys Prior

Jeffreys（1946）はパラメータ θ に対する事前分布は $[\theta] \propto \sqrt{I(\theta)}$ と定義することを提案した。これは、式 6.14 と、$\psi = \psi(\theta)$ という再パラメタリゼーションによって事前分布 $[\psi] \propto \sqrt{I(\psi)}$ をもたらす変数変換の定理から直感的に導かれる物である。そのため、Jeffreys の事前分布は変換に対し不変である。

尤度関数 $L(\theta|X)$ は $L(\theta|X) \propto g(\psi(\theta) - s(X))$ と表現できるのなら data-translated であると言える[9]。data-translated 尤度では、データは尤度の位置には影響するが形には影響しない。そのため、Box and Tiao （1973）は一様事前分布を $\psi(\theta)$ に適用するのは適切であると論じている。Jeffreys 事前分布を用いることは、一様事前分布をあてることが適切であるような状況でパラメタリゼーションを選択することと等価である; 尤度が data-translated であれば、この方法は一致する。

Jeffreys Prior for Binomial Success Rate

以下のようにして、二項分布に従う成功率のための Jeffreys 事前分布を計算する。第一に、尤度は $L(p|X) \propto p^X(1-p)^{n-X}$ である; そのため、対数尤度は定数に $X \log(p) + (n - X) \log(1 - p)$ を加えた物であり、スコア関数は以下のようである。

$$S(p|X) = \frac{X}{p} - \frac{n-X}{1-p} = \frac{X - np}{p(1-p)}$$

$E(X) = np$ とすると、$I(p) = \mathrm{Var}(X)/(p(1-p))^2$ となるのは明らかである; そのため Jeffreys 事

[9] このことを説明するために、平均 0、未知の分散 σ^2 に従う n 個の正規乱数について考える。σ^2 の尤度は data-translated であり $g(t) = \exp(-nx - e^{-x})$、$\psi(\sigma^2) = -\log(\sigma)$、$s(X) = \sqrt{\sum X^2}$ である。

前分布は $[p] \propto 1/\sqrt{p(1-p)} = p^{-1/2}(1-p)^{-1/2}$ である。この事前分布は直ちに $Be(1/2, 1/2)$ のベータ分布と認識される[10]。

少しの間、「変換に対する不変性」が意味するところを正確に考えることは価値のあることである。二項分布のパラメタリゼーションを成功確率 p ではなく、成功のオッズ比の自然対数 $\eta = \text{logit}(p)$ で推論を行うとする。事後分布 $[p|X]$ からランダムサンプルを得ることで推論を行えるので、各サンプル p は η に変換することができる。結果として得られる η の事後分布は p に導入されたある特定の事前分布に対応している。「無情報」という定義によって、p に対する無情報事前分布は η にとって無情報事前分布を導入しないかも知れない。Jeffreys 事前分布を無情報であることの定義として用いようとするなら、そのような問題は生じない。p に対して Jeffreys 事前分布を用い、p のサンプルを得、それらを η に変換すれば、η については Jeffreys 事前分布を η に適用したような事後のサンプルが得られる。

Jeffreys Prior for Multivariate Parameters

ほとんどの場合、モデルは複数のパラメータ $\theta = (\theta_1, \theta_2, ..., \theta_k)'$ を持つ。Jeffreys は 2 つの可能性について言及した。1 つは、一変量の場合に Jeffreys が推奨している手法に従い、$\theta_j, j \neq i$ が既知であるとしてそれぞれの θ_i について事前分布を得るものである。統合された事前分布 θ は（個別の）事前分布の積で得る。この方法は平均と分散がわからない正規的なデータについては意味があるが、多項分布のデータにおいてそれぞれの段階の確率がわからない状況では適さない、というのは多項分布においては合計が 1 にならなければならない（そのため独立ではない）からである。

もし変換不変性を保持したいのであれば、θ に関する Jeffreys 事前分布は Fisher 情報行列の平方根に比例するものである。Fisher 情報行列は i, j の要素を持つ以下の行列である。

$$-E(\frac{\partial L(\theta|X)}{\partial \theta_i \partial \theta_j})$$

1 パラメータの問題に Jeffreys 事前分布を使用するのは広く適切だと認められているが、多変量について Jeffreys 事前分布を用いることについてはまだ議論が残っている（Kass and Wasserman, 1996）

6.2.4 Summary

無情報事前分布として利用可能な分布として、一様事前分布、共役事前分布、Jeffreys 事前分布について言及してきた。これらと他の多くの解決法が、推論の仮定において主観性または恣意性という欠点をなくすために、どのようにして客観的なベイズ分析を行うのかという問いに対して提案されてきた。二項分布の成功確率パラメータの事例では、$Be(1,1)$、$Be(0,0)$、$Be(0.5, 0.5)$ という 3 つの事前分布による手法について検討した。サンプルサイズが小さい場合を除き、結果は実用的

[10] その累積分布は $0 < t < 1$ において $1/2 + (1/\pi)\sin^{-1}(2t-1)$ となるのでアークサイン分布としても知られている。

には区別不可能なほど差がないものであった。例えば 3 章において、10 羽の内 8 羽のベニアジサシが雌であった Shelter *et al.* の観察データについて議論した。3 つの事前分布による事後分布が Fig. 6.4 に示されている。非常にサンプルサイズは小さいが、結果は非常に似ている。成功率の 95% 最高事後密度区間は（0.52, 0.96)、(0.54, 0.98)、(0.57, 0.99）; $p \leq 1/2$ となる事後確率は 0.033、0.026、0.020 であった（それぞれ一様分布、Jeffreys 事前分布、Haldane 事前分布)。

Jeffreys 事前分布は時に非正則であるが、その挙動は「ほぼ常に正則な事後分布を得る」ので「somewhat marginal」と表現される（Yang and Berger 1996)。その形は通常非常に単純である。例えば、位置尺度分布は以下を満たす密度関数である。

$$f(t|\mu,\sigma) = \frac{1}{\sigma} g(\frac{t-\mu}{\sigma})$$

ここで $g(\cdot)$ は確率密度関数の baseline である; 位置および尺度パラメータについてはそれぞれ $-\infty \leq \mu \leq \infty$ および $\sigma > 0$ が既知とする。正規分布は、平均が位置パラメータ、標準偏差が尺度パラメータとして理想的な例を与えてくれる。$f(t|\mu,\sigma)$ の範囲は未知のパラメータに依存せず、Fisher 情報量が存在すれば、位置パラメータの Jeffreys 事前分布は一様分布、尺度パラメータについては σ^{-1} である。

Jeffreys 事前分布の興味深い特徴は、少なくともほとんどの場合「probability matching」な事前分布であると言うことである。Probability matching な事前分布とは、頻度論における信頼区間と同一の信用区間を生成するものである（Kass and Wasserman, 1996)。単純な説明として、平均と分散が未知の $N(\mu,\sigma^2)$ から得られたランダムサンプル $\mathbf{y} = (y_1, ..., y_n)'$ を用いて μ をベイズ推論する。もし非正則な事前分布 $[\mu,\sigma] \propto 1/\sigma$ を採用すれば、μ の事後分布は自由度 $n-1$ の t 分布に従う $\sqrt{n}(n-\bar{y})/2$ という性質を持つ。そのため、μ の事後推論の最高事後密度区間はこの問題における頻度論で通常用いる信頼区間と正確に一致する[11]。この例において、事前分布 $[\mu,\sigma] \propto 1/\sigma$ は σ が既知で μ に対する Jeffreys 事前分布と μ が既知で σ に対する Jeffreys 事前分布の積であり、Jeffreys 自身が推奨している事前分布である（Kass and Wasserman, 1996)。

6.3 Afterword

Little（2006）がベイズ推論の強みと弱みを表現する際に、「頻度論の概念は十分に正確な答えを提供しないが、ベイズでは尤度が決定されてしまえばあらゆる事前分布が異なる答えを導くので悩ましいことが多い」と嘆いている。我々は、「それらは全て正しい」と付け加えるだろう。

一般的に、事前分布を選択する際の問題には 2 つの解決方法がある。1 つは、主観的なベイズ推論を行う際に取り入れられるように、既知の意見を明示し、この事前分布を取り入れてモデルを更新し事後分布を得るというエレガントな過程を享受することである。もう 1 つは、推論の過程から

[11] ベイズの信用区間は 95% という点で長期間繰り返し得られているデータに対してよく補正されているという考えには大きな魅力がある。我々の 95% 信用区間は、適用時にパラメータがどんな値かは問題ではなく、長期間データが取られたときに真の値を含んでいるのである。残念ながら、ベイズの信用区間がよく補正されているという事例は少なく、頻度の補正は無情報事前分布の探索に使用するには制限のある基準である。

主観を除きたいと願っている人達に好まれる、指定した尤度のもとでデータそのものに語らせようとすることで情報が制限されていることを表現するような事前分布を選択することである。有限の範囲を持つパラメータについての一様分布、位置パラメータについての変則事前分布、尺度パラメータに関する $[\theta] \propto 1/\theta$ という事前分布、これら全ての利点を客観的な事前分布を指定したいのであれば考える必要がある。パラメータ θ の範囲が 0 以上であれば「あいまいな」事前分布、すなわち $U(0, L)$、ここで L は「大きい」という意味、が用いられる; または、$[\theta] \propto 1/\theta$ である非正則事前分布を近似できるような小さい ϵ を持った $Ga(\epsilon, \epsilon)$ を事前分布とすると言うことも考えられる。回帰係数やデータ全体に広がる他のパラメータについては平均 0 で分散が大きい正規分布が挙げられるだろう。

ただし、事前分布の選択について標準のもの以外考えないと言うことには注意する必要がある。適切な事前分布は常に適切な事後分布を導く; 不適当な事前分布はしばしば不適当な事後分布を導く; しかしながら（？）不適切な事前分布は不適切な事後分布を招き、それは深刻な問題である。そのような場合、不適切な事前分布に適切な近似を行うことは問題の解決にはならない、なぜなら事後分布もまた不適切な事後分布の近似であり、不安定な推論となるからだ。この不安定さはシミュレーションでははっきりとしない。

Gelman（2006）は階層モデルにおける分散に関するパラメータの推論の問題を議論している。そのモデルの 1 例は、

$$\begin{aligned} y_{ij} &\sim N(\mu + \alpha_i, \sigma_y^2) i = 1, ..., n_j; j = 1, ...j \\ \alpha_j &\sim N(0, \sigma_{alpha}^2), j = 1, ..., J \end{aligned} \tag{6.15}$$

ここで、μ と σ_y^2 の推論について十分なデータを持っているとする。もし $\alpha_1, ..., \alpha_J$ を観測したら、共役な事前分布である逆ガンマ分布 $IG(\alpha, \beta)$ を σ_α^2 の事前分布とする。この場合無情報な $IG(\alpha, \beta)$ の「事前分布」は適切な事後分布を導くだろう。さらに、式 6.15 に示されている階層モデルの σ_α^2 について、事前の知識がはっきりしないことを示すために ϵ に小さい値を用いた $IG(\epsilon, epsilon)$ という事前分布を使いたくなる。しかし Gelman（2006）は、$IG(\epsilon, \epsilon)$ という事前分布はこの階層モデルにおいて、$\epsilon \to \infty$ に収束すると σ_α^2 の事後分布は決して適切なものにならないことを示している。That is, the improper $IG(0, 0)$ prior will lead to improper proper posterior which an $IG(\epsilon, \epsilon)$ prior will approximate, leading to unstable inference. そのような問題のために、Gelman（2006）は σ の事前分布について大きな A を用いた一様分布 $U(0, A)$ を推奨している。この事前分布は $J \leq 3$ である限り $A \to \infty$ となるので適切な制限された事後分布を得る。

このような階層的なモデルの問題について、Gelman（2006）は「わずかに情報のある事前分布」という概念について論じている: 事前分布は実際に利用可能な事前情報をわざと控えめにしたものが含まれるように選ばれる。そのため、Gelman （2006）は無情報事前分布を探索するよりも主観的だが情報の少ない事前分布を使うように主張している。

事前分布の選択については多くの文献があり、特に有用なレビューは Kass and Wasserman（1996）である。我々は、「決まった手順によって選択された事前分布にまつわる問題は深刻で、簡単に終わりにできるものではない: データ数が少ない（推定するべきパラメータ数に対して）時、

「標準の」解決方法を信用するのは危険である; しかし when asymptotics take over、Jeffreys の規則やそれらの亜種は適切な選択である」という彼らの結論に賛成である。

到達点が主観または客観的なベイズ解析を行うことかによらず、事前分布を選択する「問題」の解決方法は、この章から始まった観察に見いだせる。解析部分として事前分布を常に報告するべきであり、事前分布の選択による影響を検証するための感受性分析を常に実施すべきである。これらのことが行われれば、事前分布についてさらに心配することは単に蚊を引っ張るようなものである（たいしたことはないと言うことが言いたい？）。