

ベイズ統計学の原理と応用（その3）

美 添 泰 人

1 はじめに

筆者は、[2, 3] において Pratt [18] および Savage [19] に沿ってベイズ統計学の原理が再構築された過程をみてきた。Savage についてはもう少し紹介する予定である。一方、もう少し基礎的な内容を踏まえた解説の必要性を感じてきたため、今回は、*International Encyclopedia of Statistical Science* [16] も参照しながら、ベイズ統計学の構造から比較的最近の話題までを扱うことにする。ベイズ統計学を扱った最近の著作は計算手法の解説が中心で、原理的な視点が欠落しているものがほとんどである中では Ghosh *et al.* [11] はバランスが取れた記述が多く、参考になる。

ベイズ統計学の手法によって、推定や検定などの古典的な統計的推論の問題に関してはもちろん、高次元データの分析という、最近では一般的になりつつある新たな問題についても明確な解答を与えることができる。その他、母数に関する部分的な知識、たとえば $a < \theta < b$ のような領域の制約や $\theta_1 \neq \theta_2$ という母数間の関係などを分析に取り入れることも容易である。

さらに経営に関する主題に関してはハーバードビジネススクールで展開されたような手法で、意思決定者 (Decision maker, DM) の持つ情報を事前分布の形で取り出し、効用関数 (損失関数) の定式化を行うなどの実績もあるが、これらの主観的判断が重要な役割を果たす問題以外では、情報のない事前分布 (客観事前分布とよぶこともある) が採用される。事前分布の選択に関しては、頑

健性などに注意が払われる。本稿でも情報のない事前分布を中心に扱うことにする。

本稿では詳細は扱わないが、古典的な決定問題では、Wald [20] による次の結果が重要である。ベイズ規則およびその極限は、完備類 (complete class) を構成する。すなわち任意の許容的な決定規則 (admissible decision rule) に対して、このクラスの中に同等以上に優れた規則が存在する。逆に、ある決定規則が許容的であれば、それはベイズ規則またはその極限となる。

ベイズの手法は意思決定者 (DM) の合理的な行動に関する公理から導かれる。その限りにおいて矛盾がないため、古典的な統計の手法において発生する矛盾を避けることができる。古典的手法の問題は、観測された実現値ではなく標本空間全体における分布を考察するため、特定のデータが与えられたときには無意味な結論を導く可能性があるという点である。区間推定および仮説検定に関しては、後にそのような例を記すように、パラドックスは際立ったものになることもある。一方で、情報のない事前分布を用いるベイズの手法は古典的な手法を部分的に正当化するために有用であることも、後に例を示す。ただしベイズの手法の最大の利点は、そこに論理的な矛盾がないことが保証されるため、結論の解釈が容易だということである。

このような利点にも関わらず、ベイズの手法が驚異的な発展を始めるまでは、21世紀を待たなければならなかった。その大きな理由は、初等的な問題以外では事後分布を解析的に評価することが困難だったものが、コンピュータの能力の圧倒的な向上と、シミュレーション手法、特に MCMC (Markov chain Monte Carlo) の開発によって、複雑な高次元の問題が実際に解けるようになったことである。Gelfand-Smith [10] は、今やこの分野の古典的文献となっている。

2 ベイズ推論と決定論

ベイズ統計学における決定論についてはさまざまな表現があるが、以下では [4, 9, 11] に準じて、推定と検定に関わるベイズ統計学の手法を中心に扱う。

2.1 主観確率と標本理論の確率

確率にはさまざまな意味があり、歴史的には賭けなど、不確実性下の意思決定における個人的な評価と特定の事象の発生割合についての予測の両方に関わってきた。頻度論的、または標本理論の立場では、特定のコインが表になる確率が $1/2$ と言うとき、多数回のコイン投げでは約半数が表になることを意味する。一方、コインの表裏に賭けるときに、どちらに賭けても違いはないと考えるという意味で、主観的な確率を $1/2$ とするという、主観的な定義もある。主観確率に関しては、長い間、理論的な根拠がないことが指摘されてきたが、20 世紀の半ばに、ようやく Savage などによって論理整合性が確立された。この点については、[2, 3] で解説しているとおり、筆者は Savage 以来の、合理的な意思決定者に関する公理体系から導かれる主観確率の立場を支持している。なお Jeffreys などは科学的仮説に適用される第 3 の種類の確率が存在する可能性を指摘しているが、本稿では扱わない。

ベイズ統計学では母数と観測前のデータはいずれも未知の量として確率分布を想定する。データについては母数 θ の下で確率モデル $f(x|\theta)$ を想定する。過去の経験を含めた判断を反映するという意味では、モデルの設定は主観的である。母数 θ の事前分布は、意思決定者（DM）の主観的な判断である。しかし、知識や経験が乏しい場合には、多数の人々に同意してもらえるような一般的な弱い事前分布を想定することも多い。このような事前分布を用いる場合、客観ベイズとよぶことが増えてきているが、これは誤解を招く表現である。弱い事前分布が適当であるという判断の主体は DM であり、あくまでも主観確率である。

最近では原理的な議論を扱う著作は多くないが、歴史的には Savage [19], Jeffreys [12], DeGroot [9], Box and Tiao [8], Berger [4] のような優れた著作がある。

2.2 ベイズ推論

未知の母数 θ に関する推論はデータに関するモデル $f(x|\theta)$ を通じて行われ、

あわせて θ に関する知識を反映する事前分布が想定される。経営に関する意思決定の分野では、DM は豊富な知識を持っていることも少なくないため、その知識を反映するように事前分布を定めることができるが、一般的な問題では、母数に関する知識は必ずしも強くない。そのような場合には、情報がほとんどないことを表現する事前分布を採用する。上記のように、後者を客観事前分布 (objective prior) とよぶことがあるが、同意できない名称である。

ベイズ統計学では事前分布 $\pi(\theta)$ とデータ x が与えられたとき、ベイズの公式によって事後分布 $\pi(\theta|x)$ を評価する。 x が連続的な確率変数であれば

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_{\Theta} \pi(\theta) f(x|\theta) d\theta} \quad (2.1)$$

となる。この式の分子は (θ, x) の同時分布、分母は x の周辺分布である。もし θ が離散的な母数であれば、(2.1) 式の積分は和で置き換える。この表記では θ は確率変数とその値のどちらに対しても用いるので、確率論のような大文字の使用法は一貫性がない。通常、ベイズ統計学では確率変数とその実現値をいずれも x と表示することが多いが、必要があれば X や $\tilde{x}$ (x -tilde) という記号を用いる。

事後分布 $\pi(\theta|x)$ は、データを通じて得られる θ に関する不確実性を表現するもので、形式的には x が与えられたときの θ の条件付き分布である。 $\pi(\theta)$ から $\pi(\theta|x)$ への推移は、データからの学習そのものである。ベイズ統計では事後分布にすべての情報が集約されているから、分析結果の報告としては事後分布が基本だが、分布の特性値、たとえば事後平均

$$E(\theta|x) = \int_{-\infty}^{\infty} \theta \pi(\theta|x) d\theta$$

や事後分散

$$\text{var}(\theta|x) = E[(\theta - E(\theta|x))^2|x] = \int_{-\infty}^{\infty} [\theta - E(\theta|x)]^2 \pi(\theta|x) d\theta$$

またはその平方根である事後標準偏差も有用である。究極的には、DM は事後分布を用いて、推定や検定のような問題を体系的に扱うことができる。推定に関しては上記の要約統計量、検定に関しては仮説のオッズを報告することが考

えられる。

例 1 簡単な例として、分散 σ^2 が既知の場合の正規分布 $N(\mu, \sigma^2)$ における平均 μ の推定を取り上げる。データ $x_1, \dots, x_n$ は独立同一分布 (i.i.d.) とする。扱いやすい事前分布として共役分布があり、この場合は正規分布 $\mu \sim N(\eta, \tau^2)$ を採用する。事前分布の分散 τ^2 は平均 η に関する確信の程度を表現する。

このとき、事後分布は次の平均と分散をもつ正規分布となることが示される。

$$E(\mu | x) = \frac{\eta/\tau^2 + n\bar{x}/\sigma^2}{1/\tau^2 + n/\sigma^2}, \quad \text{var}(\mu | x) = \frac{1}{1/\tau^2 + n/\sigma^2} = \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2} \quad (2.2)$$

すなわち、事後分布の平均は事前の推測 η と観測値 $\bar{x}$ との加重平均であり、ちらばりを分散の逆数である精度という概念で表現すると、事後分布の精度は事前分布の精度 $1/\tau^2$ と標本の精度 n/σ^2 の和になる。

古典的な Jeffreys [12] では、 τ^2 を σ^2 と比較することを提案していて、 $\tau^2 = \sigma^2/m$ とすることは、 η に関する事前情報は観測値 m 個分に相当することを意味する。このときは

$$E(\mu | x) = \frac{m\eta + n\bar{x}}{m + n}, \quad \text{var}(\mu | x) = \frac{\sigma^2}{m + n}$$

である。特に τ^2 が大きい、 n が大きく十分なデータがある場合には、事後平均は MLE $\bar{x}$ に近い。 $\pi(\mu)$ と $\pi(\mu | x)$ を比較すれば、データを通じた学習の効果を計ることができる。 ■

例 2 未知の比率 θ に関して、箱の中の N 個の球のうち、赤が $N\theta$ 個、黒が $N(1 - \theta)$ 個とする。毎回もとに戻しながら n 個を取り出す実験を考えると、観測値 $x_1, \dots, x_n$ は赤のとき 1、黒のとき 0 という値を取る、 $B(1, \theta)$ からの無作為標本である。 θ の事前分布としては、共役分布であるベータ分布 $Be(\alpha, \beta)$ を想定する。

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad 0 \leq \theta \leq 1; \alpha > 0, \beta > 0. \quad (2.3)$$

この想定では、 θ は $0, 1/N, 2/N, \dots, 1$ という値の他に区間 $[0, 1]$ 上のすべての値を取る可能性を認めているが近似として用いる。事前分布の平均は $\alpha/(\alpha + \beta)$,

分散は $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$ である。ベイズの公式から事後分布は

$$\pi(\theta|x) = \{m(x)\}^{-1} \theta^{\alpha+r-1} (1-\theta)^{\beta+(n-r)-1} \quad (2.4)$$

となる。ここで $r = \sum_{i=1}^n x_i$ は赤の数、 $m(x)$ はベイズの公式における分母である。(2.3) と比較すると、事後分布は $Be(\alpha+r, \beta+n-r)$ であり、その平均と分散は次のようになる。

$$E(\theta|x) = \frac{\alpha+r}{\alpha+\beta+n}, \quad \text{var}(\theta|x) = \frac{(\alpha+r)(\beta+n-r)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

これから、 n が大きい場合は事後分散は小さく、事後平均は MLE である $\hat{\theta} = r/n$ に近づき、事後分布は $\hat{\theta}$ の周りに集中することが読み取れる。正規分布の場合と同様に、事後平均は事前分布の平均 $E(\theta) = \alpha/(\alpha+\beta)$ と観測値の平均 r/n の加重平均として

$$E(\theta|x) = \frac{\alpha+\beta}{\alpha+\beta+n} E(\theta) + \frac{n}{\alpha+\beta+n} \frac{r}{n}$$

と表される。 ■

科学における基本的な問題に予測があり、自然な予測値として $E(\theta|x)$ を用いることができる。広く利用される事前分布と対応する事後平均 $E(\theta|x)$ は次のとおりである。まず一様分布 $U(0, 1)$ は $\alpha = \beta = 1$ としたベータ分布だから、 $E(\theta|x) = (\sum_1^n x_i + 1)/(n+2)$ である。これは Laplace と Bayes が用いた方法であるが、現在では主流とはいえない。Jeffreys の事前分布は $\alpha = \beta = \frac{1}{2}$ とするもので、このときは $E(\theta|x) = (\sum_1^n x_i + \frac{1}{2})/(n+1)$ である。これは Bernardo [6] が参照事前分布 (reference prior) とよぶもので、1 次元の母数 θ に対して広く利用されている。 $\alpha = \beta = 0$ とするとベータ分布の密度関数の積分は発散し、このような事前分布は improper prior とよばれる。ここで improper とは $\pi(\theta) \geq 0$ で積分が

$$\int_{\Theta} \pi(\theta) d\theta = \infty$$

と発散する分布を指す。improper prior に確立した訳語はないが、本稿ではとりあえず「非真」分布とよぶ。非真分布の中には情報のない分布 (non-informative prior) の一部も含まれる。

ベイズの公式を適用すると, $\sum x_i = 0$ または n とならないかぎり, 事後分布は積分が 1 となる真の分布 (proper distribution) となり, 事後平均は MLE と一致する。improper な事前分布が利用できるためには事後分布が proper となる必要があり, そのときは対応する事後分布を通じて事前分布の性質を理解することができる。

例 3 $x_1, \dots, x_n$ を $N(\mu, \sigma^2)$ からの無作為標本とする。簡単のために σ^2 は既知とする。ここで $H_0: \mu \leq \mu_0$ を $H_1: \mu > \mu_0$ に対して検定することを考える。もし μ が新薬による血糖値の降下を表すなら, μ_0 は標準薬に対応する。そこで, 事後分布 $\pi(\mu | x)$ を使って仮説の事後確率を評価する。

$$P\{H_0 | x\} = \int_{-\infty}^{\mu_0} \pi(\mu | x) d\mu, \quad P\{H_1 | x\} = \int_{\mu_0}^{\infty} \pi(\mu | x) d\mu = 1 - P\{H_0 | x\}$$

分析としてはこれらの確率を提示してもいいし, 片方の確率が十分に大きければ対応する仮説を選択してもよい。具体的には事前分布を $\mu \sim N(\eta, \tau^2)$ とすると, 事後分布は (2.2) の平均と分散をもつ正規分布だから

$$z = \frac{\mu_0 - (\frac{1}{\tau^2}\eta + \frac{n}{\sigma^2}\bar{x}) / (\frac{1}{\tau^2} + \frac{n}{\sigma^2})}{\{\frac{\sigma^2}{n}\tau^2 / (\frac{\sigma^2}{n} + \tau^2)\}^{\frac{1}{2}}}$$

として $\pi(\mu \leq \mu_0 | x) = \Phi(z)$ および $\pi(\mu > \mu_0 | x) = 1 - \Phi(z)$ となる。ここで Φ は標準正規分布の分布関数である。通常の選択は $\tau \rightarrow \infty$ とした場合であり, (improper な) 一様分布 $\pi(\mu) = c$ ($-\infty < \mu < \infty$) を採用した場合と同じく, $z = \sqrt{n}(\mu_0 - \bar{x})/\sigma$ となる。事後オッズ $P\{H_1 | x\}/P\{H_0 | x\}$ が 19:1 以上のときに H_0 を棄却するなら, これは H_0 の事後確率が 0.05 以下であることに対応し, その条件は $z \leq -1.64$, 書き換えると $\bar{x} \geq \mu_0 + 1.64 \frac{\sigma}{\sqrt{n}}$ である。これは伝統的な統計学における有意水準 $\alpha = 0.05$ の仮説検定と一致する。 ■

しかし, 単純仮説 $H_0: \mu = \mu_0$ (最近は精密仮説 sharp null hypothesis というより広い意味の用語を用いることが多い) を, 対立仮説 $H_1: \mu \neq \mu_0$ または $H_1: \mu > \mu_0$ に対して検定しようとする, 1 点 $\mu = \mu_0$ に正の確率を与えるためには違う方法で事前分布を想定しなければならないし, その場合は伝統的な統

計学と異なる結論を導く。

2.3 古典的統計学におけるパラドックス

古典的な手法は標本理論 (classical or frequentist) とよばれることもあるが、その理由は、標本空間において発生する可能性があるデータ全体に関する指標を利用することにある。実際、不偏性、検定における過誤の確率、信頼区間と信頼係数などの指標は、標本空間全体における積分や和によって導出される。以下にわかりやすい例を示す。

例4 不自然な信頼区間 伝統的な統計学では、与えられた観測値ではなく標本分布にもとづく議論を用いるため、おかしい結果を得ることがある。

(1) θ を母数として、 x_1, x_2 を $\Pr(x = \theta - \frac{1}{2}) = \Pr(x = \theta + \frac{1}{2}) = \frac{1}{2}$ という分布からの無作為標本とする。ここで (1 点からなる) 信頼区間 $\{\hat{\theta}\}$ を次のように構成する。 $x_1 \neq x_2$ のとき $\hat{\theta} = \bar{x} = \frac{x_1 + x_2}{2}$, $x_1 = x_2$ のとき $\hat{\theta} = \bar{x} + \frac{1}{2}$ 。このとき、 $\hat{\theta} = \theta$ となる確率は 75% だから、 $\{\hat{\theta}\}$ は 75% 信頼区間である。しかし、 $x_1 \neq x_2$ を観測したときには $\Pr(\hat{\theta} = \theta) = 1$, $x_1 = x_2$ を観測したときには $\Pr(\hat{\theta} = \theta) = \frac{1}{2}$ となって 75% という数値は出てこない。一方、ベイズ統計では次のようになる。事前分布を $\pi(\theta) \propto c$ とすると事後分布は $\pi(\theta | x_1, x_2) \propto p(x_1, x_2 | \theta)$ となり、これから $\pi(\theta = \bar{x} | x_1 \neq x_2) = 1, \pi(\theta = \bar{x} | x_1 = x_2) = \frac{1}{2}$ が導かれる。

(2) 類似の例として x_1, x_2 の分布を区間 $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ 上の一様分布とする。 $I(\cdot)$ を指示関数とすると、 $p(x_1, x_2 | \theta) = I(\theta - \frac{1}{2} < x_1 < \theta + \frac{1}{2})I(\theta - \frac{1}{2} < x_2 < \theta + \frac{1}{2}) = I(\theta - \frac{1}{2} < x_1, x_2 < \theta + \frac{1}{2})$ と表される。簡単のために $y = \bar{x}$ と書くと、 y の分布は θ を中心にする対称な分布であり、 $\theta - \frac{1}{2} < y < \theta$ なら $\Pr(\tilde{y} < y) = [1 + 2(y - \theta)]^2 / 2$ である。したがって $\alpha < \frac{1}{2}$ なら $\Pr(\tilde{y} < y) = \frac{\alpha}{2}$ の解は $y = \theta - \frac{1 - \sqrt{\alpha}}{2}$ であり、 $\alpha = 0.25$ のときは $y = \theta - 0.25$ となるから、 $y \pm 0.25$ は信頼係数 50% の信頼区間である。しかし、具体的に観測値として $x_1 = 0.01, x_2 = 0.99$ が与えられたときは、 $y = 0.5$ から大きく離れた θ はあり得ない。一方、ベイズ統計では次のようになる。事前分布を $\pi(\theta) \propto c$ とすると、事後分布は

$\pi(\theta | x_1, x_2) \propto p(x_1, x_2 | \theta) = I(\max x_i - \frac{1}{2} < \theta < \min x_i + \frac{1}{2})$, すなわち一様分布となる。上記の観測値に対しては, θ の範囲は $(0.99 - 0.5, 0.01 + 0.5) = (0.49, 0.51)$ であり, 信用区間を構成することは容易である。 ■

これらの例は広く知られていて, Lehmann [13](Chapter 10, Problems 27 and 28), Pratt [17], Berger and Wolpert [5] などでも取り上げている。特に Fisher はこのような現象に対処するために「条件性の原理」を示している。ある統計量 $a(x)$ の分布が母数 θ に無関係であるとき, $a(x)$ を補助統計量 (ancillary statistic) とよび, 推論に際しては a に関する条件付きで議論するのが条件性の原理 (conditionality principle) である。

例4の(1)では n が, (2)では $x_1 - x_2$ が補助統計量であり, この例ではベイズの手法と一致する合理的な解が得られる。

ところで, 古典的な統計学における最大の欠点は, 一貫した原理が存在しないことである。条件性の「原理」にしても, 一般的な原理とはいえないことは次の議論からわかる。もうひとつの原理に, 十分統計量 $t(x)$ が存在するときは推論には $t(x)$ だけにとづいて行うという「弱い尤度原理」があり, これは十分統計量を用いる議論を支持するため, 広く受け入れられている。これらふたつの原理をあわせたものは, もし二つのデータで尤度関数が一致すれば同じ結論を導くという「強い尤度原理」と同等であることが Birnbaum [7] によって示されている。しかし, 強い尤度原理によれば推論は停止規則に依存しないため, 頻度論者はこの原理を受け入れない。例えば, n 回コインを投げて r 回が表とすると, 表の出る確率 θ の不偏推定量 $\hat{\theta}$ は停止規則に依存する。 n を固定した実験なら $r \sim B(n, \theta)$ だから $\hat{\theta} = r/n$ である。これに対して r を固定した実験なら $n \sim NB(r, \theta)$ だから $\hat{\theta} = (r-1)/(n-1)$ となり, 異なる結果を導くが, 尤度はどちらの場合も $\theta^r (1-\theta)^{n-r}$ と等しい。ベイズの手法では事後分布の評価には尤度関数が利用されるため, 強い尤度原理を満たしているが, 多くの古典的な手法ではこの原理に反しているため, 矛盾が生じるといえる。[7] の解説は美添 [1] に記したが, この結果が衝撃的であったことは, [7] のコメントから理

解できる。そのうち、Savage と Pratt のコメントは特に興味深い。Pratt は古典的な統計学の欠陥を明らかにする面白い例を指摘している。また Savage は、この性質があるから自分はベジアン立場を選択したと述べている。つまり、1954 年に出版された歴史的な著作 [19] を執筆した時点では、すでにこの結果を知っていたことになる。

2.4 ベイジアン決定理論の基本的な構造

統計的な推測は決定理論の枠組みで整理することが一般的である。推測の対象となる母数 θ の含まれる母数空間 $\theta \in \Theta$ は多くの場合 p 次元ユークリッド空間 R^p である。古典的な統計学では、標本空間を $x \in \mathcal{X}$ として、データ $\mathbf{x} = \{x_1, \dots, x_n\}$ とその分布 $f(\mathbf{x} | \theta)$ を想定する。さらに、決定 (decision) d (決定空間は $d \in D$) と母数によって定められる損失関数 $L(\theta, d) \geq 0$ が設定される。その上で、データの関数として決定を選択する決定関数 (または決定規則) $\delta(x)$ を考える。 $\delta(x) = d$ のときに発生する損失は $L(\theta, \delta(x)) = L(\theta, d)$ である。推定と検定はこの枠組みによって表現される。

母数 θ の実数値関数である $\tau(\theta)$ を推定する問題では、 $D = R$ であり、 $\delta(x)$ は $\tau(\theta)$ の推定量である。損失関数は $L(\theta, d) = (d - \tau(\theta))^2$ とすることが多い。

仮説 $H_0 : \theta = \theta_0$ を対立仮説 $H_1 : \theta \neq \theta_0$ に対して検定する問題では、 d_0, d_1 を H_0, H_1 を受け入れる決定として決定空間は $D = \{d_0, d_1\}$ となる。また、損失関数は H_j を満たす θ の集合を Θ_j とすれば

$$L(\theta, d_j) = 0 \quad (\theta \in \Theta_j \text{ のとき}), \quad L(\theta, d_j) = 1 \quad (\theta \notin \Theta_j \text{ のとき}) \quad (2.5)$$

と表すことができる。決定関数 $\delta(x)$ は d_0 または d_1 のいずれかの値を取る。

古典的な統計学では $\delta(x)$ の評価には損失関数の期待値であるリスク

$$R(\theta, \delta) = E[L(\theta, \delta(x)) | \theta] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x | \theta) dx$$

が用いられる。 $R(\theta, \delta)$ は θ の関数だから、最適な決定は θ に依存する。

$\tau(\theta)$ の推定問題においては δ は推定量であり、損失関数が $L(\theta, d) = (d - \tau(\theta))^2$ なら、 $R(\theta, \delta) = E[(\tau(\theta) - \delta(x))^2 | \theta]$ は平均 2 乗誤差 (MSE, mean squared error) で

ある。検定の問題で、 δ を H_0 の棄却域に対する指標関数とすれば、 $R(\theta, \delta)$ は、 $\theta \in \Theta_0$ または $\theta \in \Theta_1$ に応じて第1種過誤の確率または検出力である。

ベイズ統計学では θ は確率変数であり、データを観測する前には $\pi(\theta)$ による決定 d の期待損失は

$$E[L(\theta, d)] = \int_{\Theta} L(\theta, d) \pi(\theta) d\theta$$

である。またデータ x を観測した後では、 $\pi(\theta|x)$ による事後期待損失は $E[L(\theta, d)|x]$ となり、これも決定 d の関数である。最適な決定関数は与えられた x に対して $E[L(\theta, \delta(x))|x]$ を最小にする $\delta(x)$ である。この問題に関してはふたつの導出方法がある。

- (1) 与えられた x に対して、事後期待損失 $E[L(\theta, d)|x]$ を最小にする決定 d を選ぶ。
- (2) 最初の段階でリスクの $\pi(\theta)$ による期待値

$$E[R(\theta, \delta(x))] = \int_{\Theta} E[L(\theta, \delta(x))|\theta] \pi(\theta) d\theta$$

を最小にするように $\delta(x)$ を選ぶ。

もちろん、後の問題の方が難しいが、これらの方法は同じ結果を与える。すなわち $E[R(\theta, \delta(x))]$ は次のように変形できる。

$$\begin{aligned} \int_{\Theta} \left(\int_x L(\theta, \delta(x)) f(x|\theta) dx \right) \pi(\theta) d\theta &= \iint_{\Theta \times X} L(\theta, \delta(x)) f(x, \theta) dx d\theta \\ &= \int_x \left(\int_{\Theta} L(\theta, \delta(x)) \pi(\theta|x) d\theta \right) f(x) dx \end{aligned}$$

ここで $f(x)$ はベイズの定理の分母に現れる x の周辺分布であり、 $f(x|\theta)\pi(\theta) = f(x, \theta) = \pi(\theta|x)f(x)$ となることを用いている。最後の式において $()$ の内では x は一定だから、事後分布による損失の期待値を最小にする d を選べば、最適な関数 $\delta(x) = d$ が求められる。

ところで、事後期待損失 $E[L(\theta, d)|x]$ の概念は重要であるが、必ずしも用語として統一されていない。たとえば DeGroot [9] では $\rho(\pi, d) = \int_{\Omega} L(\omega, d) dP(\omega)$ を “expected loss or risk”, $\rho^*(\pi) = \inf_{d \in D} \rho(\pi, d)$ を “Bayes risk”, $L(\omega, d) = -U[\sigma(\omega, d)]$

を“Loss”とよんでいる。母数に ω という記号を用いているが、損失は負の効用であり、 $\sigma(\omega, d)$ が損失関数に明示的に現れる報酬 (award) であることがわかれば、意味は明確であろう。Lindley [15] では Wald に倣って $R(\delta, \theta) = \int_{\mathcal{X}} L[\delta(x), \theta] p(x|\theta) dx$ を “risk function for δ ” とし、 $\int_{\Theta} U(d, \theta) p(\theta|x) d\theta = \max!$ を “posterior expected utility” とよんでいる。ベイズ統計学の原理を明確に提示した Savage [19] では公理的に導出された効用はすでに母数に関する期待値だから、事後期待効用に相当する概念に対しては “conditional utility” となる。そもそも Savage はこの段階ではベイズ統計学という表現は用いていない。

2.5 情報のない事前分布

これまでの例に示したように、 $N(\mu, \sigma^2)$ における μ の推定問題では情報のない非真 (improper) 事前分布を用いることが多いし、検定についても非真分布を用いることがある。これらの分布が利用可能なのは、事後分布を与える式の分母 $f(x)$ がすべての x に関して有限であることが前提となる。

$$\int_{\Theta} \pi(\theta) f(x|\theta) d\theta < \infty$$

このとき、事後分布 $\pi(\theta|x)$ は真の分布 (proper prior) となって、事後期待損失を評価することができるし、それを最小にする最適な決定関数を導出することも可能となる。なお、非真分布に対しては、通常、事前期待損失 $E[L(\theta, d)]$ は利用できない。

位置母数 μ と尺度母数 σ に関する一般的な非真事前分布は次のとおりである。

$$\pi_1(\mu) = c \quad (-\infty < \mu < \infty), \quad \pi_2(\sigma) = \frac{1}{\sigma} \quad (0 < \sigma < \infty)$$

これらは、次に示す真の事前分布の極限と理解することができる。

$$\pi_M(\mu) = 1/(2M) \quad (-M < \mu < M), \quad \pi_M(\sigma) = A/\sigma \quad (1/M < \sigma < M)$$

ただし $A = 1/(2 \log M)$ である。 $\pi_M(\theta)$ において $M \rightarrow \infty$ とした場合が、非真事前分布である。

2.6 推定と検定におけるベイズ統計学の手法

区間推定と片側仮説検定についてはすでに触れたが，ここでは，古典的な統計学に準じて推定と検定に関するベイズ統計学の推論を紹介する。

点推定 実数値を取る母数 θ に関しては，標準的なベイズ推定量は損失関数 $L(\theta, d) = (\theta - d)^2$ または $L(\theta, d) = |\theta - d|$ に対応する事後平均または事後メディアンであり，あわせて事後分散（あるいは標準偏差），四分位などを提示することが多い。事後分布が一山型 (unimodal) なら，事後モードも選択肢の一つになる。これは MLE に類似しているし，事前分布が一様分布であれば完全に一致する。事後モードとともに，最大事後密度 (highest posterior density, HPD) 区間，多次元の母数に対しては HPD 領域，を提示することもできる。多変量正規分布の場合には，HPD 領域は楕円の内部となることはよく知られている。

検定 もし仮説が $H_0: \theta \in \Theta_0$ と $H_1: \theta \in \Theta_1$ であり， Θ_0 と Θ_1 が，片側検定の場合のように同じ次元をもつ空間であれば，通常，事前分布として Θ_0 と Θ_1 のいずれにも正の確率を与えるものを選ぶ。これから事後確率 $P(\Theta_0|x)$, $P(\Theta_1|x)$ や事後オッズ $P(\Theta_0|x)/P(\Theta_1|x)$ を評価し，事後オッズが $1/19$ などより小さいとき（すなわち事後確率が $P(\Theta_0|x) < 0.05$ などより小さいとき）に， H_0 に反する証拠が強いと判断することができる。もし損失関数 $L(\theta, d)$ が (2.5) 式のような 0-1 型であれば，事後確率が大きい仮説を選択することになる。

しかし，この方法には基本的な問題がある。まず，非真事前分布を使うと事前確率を評価できない。また，真の事前分布であっても， Θ_0, Θ_1 の確率を適切に設定することが難しい。たとえば古典的統計学のように H_0 を棄却するためには十分に強い証拠が必要だと想定するなら， Θ_0 の事前確率を Θ_1 の事前確率より小さく設定することは合理的ではないため，双方の確率を等しく設定するような客観的または偏りのない分布は使えない。

この問題を解決するために，以下に記すような方法が利用される。この方法は，より一般的に複数のモデルを比較する問題でも利用されるものである。 Θ_0 と Θ_1 の事前確率を π_0 および $1 - \pi_0$ として， Θ_i における事前分布を $g_i(\theta)$,

$\int_{\Theta_i} g_i(\theta) d\theta = 1$ と想定する。全体の事前分布は

$$\pi(\theta) = \pi_0 g_0(\theta) I\{\theta \in \Theta_0\} + (1 - \pi_0) g_1(\theta) I\{\theta \in \Theta_1\} \quad (2.6)$$

である。このとき、 Θ_0 と Θ_1 の次元は異ってもよいが、単純仮説の問題も含まれることになる。なお、 $\Theta = R^p$ のときは R^p における Lebesgue 測度が 0 となるような仮説 Θ_0 を精密仮説 (sharp or precise null hypothesis) とよぶことが多い。この想定であれば、事後確率および事後オッズを評価することができる。

x の周辺分布 $m(x)$ は

$$\begin{aligned} m(x) &= \int_{\Theta} f(x|\theta) \pi(\theta) d\theta \\ &= \pi_0 \int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta + (1 - \pi_0) \int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta \end{aligned} \quad (2.7)$$

だから、データ x が与えられたときの事後分布は

$$\pi(\theta|x) = \frac{f(x|\theta)}{m(x)} = \begin{cases} \pi_0 f(x|\theta) g_0(\theta) / m(x) & \theta \in \Theta_0 \text{ のとき,} \\ (1 - \pi_0) f(x|\theta) g_1(\theta) / m(x) & \theta \in \Theta_1 \text{ のとき} \end{cases} \quad (2.8)$$

となる。これから $P(H_0|x) = P(\Theta_0|x)$ および $P(H_1|x) = P(\Theta_1|x) = 1 - P(H_0|x)$ に関する表現が得られる。

$$\begin{aligned} P(H_0|x) &= \frac{\pi_0}{m(x)} \int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta \\ P(H_1|x) &= \frac{1 - \pi_0}{m(x)} \int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta \end{aligned}$$

さらに π_0 に依存しない指標として、次式で定義されるベイズ因子 (Bayes factor) も用いられる。

$$BF_{01} = \frac{\int_{\Theta_0} f(x|\theta) g_0(\theta) d\theta}{\int_{\Theta_1} f(x|\theta) g_1(\theta) d\theta} \quad (2.9)$$

H_0 の H_1 に対する事後オッズは $\frac{P(H_0|x)}{P(H_1|x)} = \left(\frac{\pi_0}{1 - \pi_0} \right) BF_{01}$ であり、特に $\pi_0 = \frac{1}{2}$ なら BF_{01} と一致する。 BF_{01} は π_0 に依存しない指標として、小さいほど H_0 に反する証拠が強いと解釈される。

片側仮説におけるように母数空間全体に対して連続型的事前分布が合理的に

設定できる場合なら，それを $\pi(\theta) = \pi_0 g_0(\theta) I\{\theta \in \Theta_0\} + (1 - \pi_0) g_1(\theta) I\{\theta \in \Theta_1\}$ のような形で表現する必要はない。しかしながら，1 点仮説（精密仮説）の検定に関しては，状況は大きく異なる。

精密仮説の検定 統計的品質管理のような状況で θ が製品の規格であれば， $\theta \in (\theta_0 - \delta, \theta_0 + \delta)$ となることが求められ，このときの仮説は $H_{0\delta} : |\theta - \theta_0| \leq \delta$ （区間帰無仮説）とするのが合理的であり，区間の幅 2δ も指定される。このような場合には π_0, g_0, g_1 の設定には問題はない。同様に 1 点仮説が区間仮説 $H_0 : \theta \in (\theta - \epsilon, \theta + \epsilon)$ で近似できるのであれば， π_0 は連続型分布によって区間 $(\theta - \epsilon, \theta + \epsilon)$ に与えられる確率である。

一方で，ある薬の効果の存在に関しては，何らかの効果は必ず存在するから，実際は非常に短い区間に関する仮説となるとしても，その区間の幅はわからない。このような場合には 1 点仮説（単純仮説） $H_0 : \theta = \theta_0$ と対立仮説 $H_1 : \theta \neq \theta_0$ を考えることに意味がある。また，物理的な定数に関する仮説も 1 点仮説となる。

1 点仮説（一般には精密仮説）の検定に関しては，連続型の事前分布では H_0 に正の確率を与えることができないため，1 点 θ_0 に事前確率 $\pi_0 > 0$ を付与して， $\{\theta \neq \theta_0\}$ に対しては密度関数 g_1 を用いて $\pi_1 = 1 - \pi_0$ を与える。この場合の g_0 は θ_0 に確率 1 を与える分布であり，事前分布 π は次のように離散型の部分と連続型の部分を持つ。

$$\pi(\theta) = \pi_0 I\{\theta = \theta_0\} + (1 - \pi_0) g_1(\theta) I\{\theta \neq \theta_0\} \quad (2.10)$$

ここで

$$m_1(x) = \int_{\theta \neq \theta_0} f(x|\theta) g_1(\theta) d\theta$$

とおけば，周辺分布 (2.7) は $m(x) = \pi_0 f(x|\theta_0) + (1 - \pi_0) m_1(x)$ となり，事後分布 (2.8) は次のようになる。

$$\begin{aligned} \pi(\theta_0|x) &= \frac{\pi_0 f(x|\theta_0)}{m(x)} = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + (1 - \pi_0) m_1(x)} \\ &= \left\{ 1 + \frac{(1 - \pi_0)}{\pi_0} \frac{m_1(x)}{f(x|\theta_0)} \right\}^{-1} \end{aligned} \quad (2.11)$$

事後オッズは

$$\frac{\pi(\theta_0 | x)}{1 - \pi(\theta_0 | x)} = \frac{\pi_0}{(1 - \pi_0)} \frac{f(x | \theta_0)}{m_1(x)}$$

であり, H_0 の H_1 に対するベイズ因子 BF_{01} は事後オッズを事前オッズ $\pi_0/(1-\pi_0)$ で割った $f(x | \theta_0)/m_1(x)$ となる。これを用いれば (2.11) 式は次のように表される。

$$\pi(\theta_0 | x) = \left\{ 1 + \frac{1 - \pi_0}{\pi_0} BF_{01}^{-1}(x) \right\}^{-1} \quad (2.12)$$

なおベイズ因子は $B, B(x), BF_{01}(x)$ などとも表される。

Jeffreys の議論 Jeffreys [12] は, 分散が未知の正規分布 $N(\mu, \sigma^2)$ からの無作為標本 $x_1, \dots, x_n$ にもとづく仮説 $H_0: \mu = \mu_0$ および対立仮説 $H_1: \mu \neq \mu_0$ の検定を扱っている。以下, 簡単のために $\mu_0 = 0$ とする。この場合, 分散 σ^2 は H_0 および H_1 に対応するモデルに共通であり, μ は H_1 だけに現れる。さらに μ と σ^2 は Fisher 情報量が対角行列となるという意味で直交している。

Jeffreys は σ^2 に情報のない事前分布, μ に真の事前分布を用いる。すなわち, H_0 の下で σ の事前分布を $g_0(\sigma) \propto \sigma^{-1}$ とする一方, H_1 の下では σ の事前分布を $g_0(\sigma)$, σ が与えられたときの μ の条件付き分布を $g_1(\mu | \sigma) = \sigma^{-1} g(\sigma^{-1} \mu)$ と想定する。ここで $g(\cdot)$ は密度関数である。結局 H_1 の下では (μ, σ) の事前分布は $g_1(\mu, \sigma) = g_0(\sigma) \sigma^{-1} g(\sigma^{-1} \mu)$ となり, このときベイズ因子は

$$BF_{01} = \frac{\int f(\mathbf{x}, \mu = 0, \sigma^2) g_0(\sigma) d\sigma}{\iint f(\mathbf{x}, \mu, \sigma^2) g_1(\mu, \sigma) d\mu d\sigma}$$

である。もし $s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$ が有界なままで $\bar{x} \rightarrow \infty$ となれば, ベイズ因子 BF_{01} はゼロに近づくことが期待されるが, Jeffreys は g_1 がモーメントを持つなら, たとえば g が正規分布 $N(0, \tau^2)$ なら, そうはならないことを示した。筆者は, この結果は精密仮定の検定方法には十分な注意が必要であることを示していると考えている。

記号は原典とは異なるが, 以下に Jeffreys の議論を紹介する。

データの分布は

$$f(\mathbf{x} | \mu, \sigma^2) = (2\pi)^{-n/2} \sigma^{-n} \exp\{-(1/2\sigma^2)[n(\bar{x} - \mu)^2 + ns^2]\}$$

である。十分統計量 $\bar{x}, s^2$ を使えば, これらは互いに独立で $\bar{x} \sim N(\mu, \sigma^2/n)$ かつ $y = ns^2/\sigma^2 \sim \chi^2(n-1)$ であり, $f_y(y) = \frac{1}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} y^{\frac{n-1}{2}-1} e^{-\frac{y}{2}}$ と $f_{s^2}(s^2) = (n/\sigma^2) f_y(ns^2/\sigma^2)$ から

$$f(s^2) = \frac{n}{\sigma^2} \cdot \frac{1}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \left(\frac{ns^2}{\sigma^2} \right)^{\frac{n-1}{2}-1} e^{-\frac{ns^2}{2\sigma^2}} = \frac{1}{\Gamma(\frac{n-1}{2})} \left(\frac{n}{2\sigma^2} \right)^{\frac{n-1}{2}} (s^2)^{\frac{n-3}{2}} e^{-\frac{ns^2}{2\sigma^2}}$$

と表される。結局, 同時密度関数は次式となる。

$$\begin{aligned} f(\bar{x}, s^2 | \mu, \sigma^2) &= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\Gamma(\frac{n-1}{2})} \left(\frac{n}{2\sigma^2} \right)^{\frac{n-1}{2}} (s^2)^{\frac{n-3}{2}} e^{-\frac{ns^2}{2\sigma^2}} \\ &\propto \sigma^{-n} e^{-\frac{n}{2\sigma^2} \{(\bar{x}-\mu)^2 + s^2\}} \end{aligned}$$

したがって BF_{01} の分子と分母の被積分関数は次のようになる。

$$\begin{aligned} f(\mathbf{x}, \mu = 0, \sigma^2) g_0(\sigma) &\propto \sigma^{-n-1} e^{-\frac{n}{2\sigma^2} \{\bar{x}^2 + s^2\}}, \\ f(\mathbf{x}, \mu, \sigma^2) g_1(\mu, \theta) &\propto g\left(\frac{\mu}{\sigma}\right) \sigma^{-n-2} e^{-\frac{n}{2\sigma^2} \{(\bar{x}-\mu)^2 + s^2\}} \end{aligned}$$

$n=1$ ならデータから σ に関する情報を得られないため, x がゼロと離れていても, それが σ による誤差なのか μ が 0 と異なる影響なのか識別できず, 仮説の判定もできない。 $n \geq 2$ に対して s が小さく, $\bar{x} \neq 0$ であれば, それは $\mu \neq 0$ を支持するものと考えられる。そこで $s=0, \bar{x} \neq 0$ として BF_{01} を評価する。変数を $\sigma = \bar{x}/u, \mu = \sigma v = \bar{x}v/u$ と変換する。このとき $d\sigma/du = -\bar{x}/u^2$, ヤコビアンは $J = d(\mu, \sigma)/d(u, v) = -\bar{x}^2/u^3$ であり, BF_{01} の分子

$$\int_0^\infty \left(\frac{u}{\bar{x}} \right)^n e^{-\frac{nu^2}{2}} \frac{du}{u} \quad (2.13)$$

は $n \geq 1$ で収束する。分母の積分

$$\int_0^\infty \frac{du}{u} \int_{-\infty}^\infty \left(\frac{u}{\bar{x}} \right)^n g(v) e^{-\frac{n(u-v)^2}{2}} dv \quad (2.14)$$

は, $n=1$ かつ g が偶関数で $g(-v) = g(v)$ となるときは,

$$\int_0^\infty \int_0^\infty \frac{1}{\bar{x}} g(v) e^{-\frac{1}{2}(v-u)^2} du dv = \frac{\sqrt{2\pi}}{\bar{x}} \int_0^\infty g(v) dv = \frac{1}{2} \frac{\sqrt{2\pi}}{\bar{x}}$$

となる。 $n=1$ なら分子も $\frac{1}{2} \frac{\sqrt{2\pi}}{\bar{x}}$ と一致するから $BF_{01} = 1$ である。これは上述

の直感的な結果であり, H_0 と H_1 の判定ができない。

そこで $n \geq 2$ のときを考える。 $s = 0, \bar{x} \neq 0$ のときに $BF_{01} = 0$ となる条件は (2.14) が発散することである。 $v > 0$ で大きいときは,

$$\int_0^\infty u^n e^{-\frac{n(u-v)^2}{2}} \frac{du}{u} \sim Nv^{n-1} \quad (2.15)$$

となる。このことは $t = u - v$ とおくと

$$\begin{aligned} \int_{-v}^\infty (t+v)^{n-1} e^{-\frac{nt^2}{2}} dt &= v^{n-1} \int_{-v}^\infty \left(1 + \frac{t}{v}\right)^{n-1} e^{-\frac{nt^2}{2}} dt \\ &\sim v^{n-1} \int_{-\infty}^\infty e^{-\frac{nt^2}{2}} dt \sim Nv^{n-1} \end{aligned}$$

から確かめられる。一方 $v < 0$ なら, $u > 0$ の範囲で $-(u-v)^2 < -u^2$ だから

$$\int_0^\infty u^{n-1} e^{-\frac{n(u-v)^2}{2}} du < \int_0^\infty u^{n-1} e^{-\frac{nu^2}{2}} du$$

となるが, 右辺は明らかに有界である。以上から, (2.14) が発散する条件は

$$\int_0^\infty v^{n-1} g(v) dv$$

が発散することである。特に, g が正規分布の場合, この積分は収束するから, $\bar{x}$ がどれほど大きくなっても $BF_{01} \rightarrow 0$ とはならない。 $n > 1$ で積分が発散する最も簡単な g として Jeffreys は Cauchy 分布

$$g(v) = \frac{1}{\pi(1+v^2)}$$

を提案している。

信用区間 θ に関するベイズ信用区間は古典的な推測における信頼区間と類似しているが異なる概念であり, その定義は次のとおりである。確率 $0 < \alpha < 1$ に対してある部分集合 $C \subset \Theta$ が $P\{C|x\} = \int_C \pi(\theta|x) d\theta = 1 - \alpha$ を満たすとき C を $100(1 - \alpha)\%$ の信頼領域 (credible region) とよぶ。

事後分布が一山型であれば C は区間である。事後分布 $\pi(\theta|x)$ が連続型であれば, $\alpha_1 + \alpha_2 = \alpha$ として, α_1 および $(1 - \alpha_2)$ 分位点を $\theta^{(1)}, \theta^{(2)}$ とするとき, 区間 $C = \{\theta | \theta^{(1)} < \theta < \theta^{(2)}\}$ について $P(C|x) = 1 - \alpha$ がなりたつ。もし, 両側の確率を等しくするなら $\alpha_1 = \alpha_2 = \alpha/2$ である。古典的な統計学における信頼係

数は真の θ を含む確率を表さないが、信用区間は $1 - \alpha$ は C が真の θ を含む確率である。

ところで、正規分布や二項分布などの母数の推定という簡単な問題では、情報のない事前分布を用いた場合の信用区間は古典的な信頼区間に近い。これが、信頼区間を用いた応用において致命的な失敗が多くないことの説明になる。

両側の確率を等しくすると、対称な分布でない限り最短の信用区間とはならない。そこで最大事後密度 (Highest Posterior Density, HPD) 領域が用いられる。もし事後分布 $\pi(\theta|x)$ が一山分布であれば、 $C = \{\theta : \pi(\theta|x) \geq c\}$ として、 c を $P(C|x) = 1 - \alpha$ が成立するように定めれば、 C が $(1 - \alpha)$ HPD 区間となることは明らかである。特に $\pi(\theta|x)$ が正規分布なら、事後平均 $m = E(\theta|x)$ と事後標準偏差 $s = \sqrt{\text{var}(\theta|x)}$ を用いて、 $(1 - \alpha)$ HPD 区間は $m \pm z_{\alpha/2}s$ で与えられる。

精密仮説の検定と信用区間 古典的な統計学では、有意水準 α の両側検定で仮説 $\mu = \mu_0$ が棄却されないことと、 μ_0 が信頼係数 $(1 - \alpha)$ の信頼区間に含まれることは同等である。たとえば、帰無仮説を $H_0 : \mu = \mu_0$ 、対立仮説を $H_1 : \mu \neq \mu_0$ とする両側検定で検定統計量が正規分布にしたがうときは、有意水準 α の受容域 $|\sqrt{n}(\bar{x} - \mu_0)/\sigma| < z_{\alpha/2}$ を変形すると $\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu_0 < \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}$ となり、 μ_0 は信頼区間に含まれる。この関係は片側対立仮説に関してもなりたつし、一般の検定統計量についてもなりたつ。

これに準じて、ベイズ統計学でも 1 点仮説 $H_0 : \theta = \theta_0$ について、信用区間に θ_0 が含まれるときに仮説を受容するという方法がある。これは Lindley の初期の著作 [14] でも用いられた。ところで、ときに、この方法と 1 点仮説の検定の結果を比較して、それらが大きく異なるという批判があるが、それは手法の欠陥ではなく、当然の結果である。信用区間を構成するために、母数空間全体に事前密度関数 $\pi(\theta)$ を想定すると、1 点 θ_0 の確率はゼロである。これに対して、すでに取り上げた 1 点仮説の検定では θ_0 に正の確率を与えることから明らかのように、ふたつの接近方法では異なる事前分布を用いているのであり、全く批判するにはあたらない。

2.7 予測の問題

無作為標本 $\mathbf{x} = \{x_1, \dots, x_n\}$ の分布を $f(x|\theta)$ とする。ここで次の値 x_{n+1} を予測することを考える。2乗損失にもとづく推定量を $t = t(\mathbf{x})$ とすると、

$$\begin{aligned} E[(x_{n+1} - t)^2 | \mathbf{x}] &= E[\{(x_{n+1} - E(x_{n+1} | \mathbf{x})) - (t - E(x_{n+1} | \mathbf{x}))\}^2 | \mathbf{x}] \\ &= E\{(x_{n+1} - E(x_{n+1} | \mathbf{x}))^2 | \mathbf{x}\} + (t - E(x_{n+1} | \mathbf{x}))^2 \end{aligned}$$

と表され、期待損失は $t = E(x_{n+1} | \mathbf{x})$ で最小値を取る。 θ を固定したときは x_{n+1} と $\mathbf{x}$ は独立だから $\pi(x_{n+1} | \mathbf{x}, \theta) = f(x_{n+1} | \theta)$ であり、予測分布は

$$\pi(x_{n+1} | \mathbf{x}) = \int_{\Theta} \pi(x_{n+1} | \mathbf{x}, \theta) \pi(\theta | \mathbf{x}) d\theta = \int_{\Theta} f(x_{n+1} | \theta) \pi(\theta | \mathbf{x}) d\theta$$

となる。 x の期待値を $\mu(\theta) = \int_{-\infty}^{\infty} x f(x | \theta) dx$ と表すと、

$$E(x_{n+1} | \mathbf{x}) = E[\mu(\theta) | \mathbf{x}] = \int_{\Theta} \mu(\theta) \pi(\theta | \mathbf{x}) d\theta$$

だから、 θ 自身が期待値 μ であれば、予測量は μ の事後平均 $\int_{-\infty}^{\infty} \mu \pi(\mu | \mathbf{x}) d\mu$ となる。予測信用区間を構成するなら、 x_{n+1} の予測分布で分位点を適当に選ばばよい。

3 完全ベイズ有意性検定 (FBST)

この節では、[16] に収録されている「完全ベイズ有意性検定 (Full Bayesian Significance Test, FBST)」という項目について概説し、論評したい。この項目の執筆者は Carlos Pereira であり、「FBST は問題の代替的な解であり、ベイズ因子に基づいて頻度論者および標準的なベイズ検定によって提起される問題のいくつかを緩和する」と主張している。

以下では 1 次元母数 $\Theta \subset \mathcal{R}$ と (十分統計量の) 1 次元標本空間 $\mathcal{R}$ の問題に限定する。この場合、精密仮説とは $H_0: \theta = \theta_0$ の形の命題である。このとき頻度論者は少なくとも x と同じ程度に θ_0 と矛盾する標本点の集合 $C \subset \mathcal{X}$ を求めるが、ベイズアンは θ_0 以上に x と整合的な母数の集合 $T \subset \Theta$ を求める。このよ

うな理解は、標本理論とベイズ理論との間の部分的な双対性と解釈される。 H を支持する証拠は、頻度論者にとっては通常の p 値 $pv = P\{x \in C | \theta_0\}$ が大きいことであるが、Pereira は、 $ev = 1 - \underline{ev} = 1 - P\{\theta \in T | x\}$ を用いて、 ev が大きいほど H を支持する証拠が強いと解釈する。一般の問題では、母数および標本空間の次元による複雑さはなく、 ev を計算するためには事後分布が評価できれば十分である。なお、Pereira は「データと帰無仮説との間に矛盾がないことの評価に母数に関する事前の評価を含めてはならないと感じる場合には、正規化された尤度で事後分布を置き換えることができる」と記しているが、それではベ이지アン立場とは認められない。また「 ev の計算には数値的な最適化と数値積分が必要になるにしても、漸近的な方法は不要である」という主張も説得力に欠ける。

pv と ev は 0 から 1 の間の値を取るという事実は「どのくらい小さければ有意か」という問への簡単な解答としてはあまり役に立たない。 p 値については、臨界値を主観的に定めることになるが、 ev によるベ이지アン有意性の評価については、ある損失関数を最小化する問題として、 ev に基づく判定は形式的なベイズ検定と解釈できる。

FBST は、精密仮説の有意性検定の必要性を想定して作られたという。精密仮説 (sharp hypothesis) の形式的な定義は次のとおりである。一般的な統計的問題で扱う空間として、母数空間 $\Theta \subset \mathcal{R}^m$ および標本空間 $\mathcal{X} \subset \mathcal{R}^k$ を考えるとき、 θ が部分多様体 Θ_H に属するという仮説の次元が Θ の次元より小さいとき、 Θ_H を精密仮説とよぶ。

H が精密仮説であれば、部分集合 Θ_H のルベーグ測度は 0 である。結果的に精密仮説を定義する集合の確率は 0 のままとなる。正の確率を割り当てて H の性質を変えるのとは対照的に、任意の $\theta \in \Theta_H$ より高い事後密度を持つ点の集合 T を考えて、 T の事後確率が大きい場合に H が棄却される。この発想は以下のように定式化される。

標準的な統計モデルを考えて、ある整数 m に対して母数を $\theta \in \Theta \subset \mathcal{R}^m$ 、 $g(\cdot)$ を Θ 上の事前確率密度、 x を観測値（スカラーまたはベクトル）、 $L_x(\cdot)$

をデータ x から得られる尤度とすると, ev を求めるには θ の事後密度関数 $g_x(\theta) = g(\theta|x) \propto g(\theta)L_x(\theta)$ のみが必要である。単純化のために, 以下では Θ_H の代わりに H と記し, 精密仮説 $H: \theta \in \Theta_H$ と, $g^* = \sup_{\theta \in H} g_x(\theta)$ および $T = \{\theta \in \Theta: g_x(\theta) > g^*\}$ を考える。このとき, 形式的な Bayesian evidence value は, 集合 T の事後確率として次の式で定義される。

$$\underline{ev} = P\{\theta \in T | x\} = \int_T g_x(\theta) d\theta$$

H を支持する証拠値 $ev = 1 - \underline{ev}$ は, 対立仮説 A (これは精密仮説ではない) を否定する証拠ではない。同様に, $\underline{ev}$ は H を否定する証拠であるが, A を支持する証拠ではない。Pereira は, $ev = 1 - \underline{ev}$ が小さいときに限って H を棄却する手順を FBST とよぶ。また ev , pv , および BF の値の間の大きな違いは, Θ が 1 次元の部分集合であれば, ほとんど生じないと主張するが, 上述のとおり, ev と BF が大きく異なる例は少なくない。

筆者の理解では, FBST は信用区間を検定に利用することに他ならない。

中間的なまとめ

美添 [3] では, 次のように記した。

Savage [19] の 3 章は, 本稿 ([3]) で紹介した内容の後, 3 章では §3.5 質的および量的な条件付確率 (Conditional probability, qualitative and quantitative), §3.6 経験を通じた確実性への接近 (The approach to certainty through experience), §3.7 事象の対称な列 (Symmetric sequences of events) と続くが, この部分は次の稿で取り上げることにしたい。

今回は, 代わりに, 基本的なベイズ統計学の手法を改めて概観することになった。Savage の紹介については, 次回以降に改めて取り上げる。

本稿の最後に, ベイズ推定量が完備類を構成することの応用として, ミニマックス推定量を構成する例を示そう。 $x_1, \dots, x_n$ を $B(1, p)$ からの無作為標本

として、平方誤差 MSE を損失関数とすると、ミニマックス推定量を求めるために、事前分布を $Be(a, b)$ とするベイズ推定量 $\hat{p} = (a+x)/(a+b+n)$ の中から候補を探すことを試みる。 $c = a+b+n$ と書くと MSE は次式で与えられる。

$$\begin{aligned} MSE &= \frac{np(1-p)}{c^2} + \left(\frac{a+np}{c} - p \right)^2 = \frac{1}{c^2} [np - np^2 + (a + (n-c)p)^2] \\ &= \frac{1}{c^2} [np - np^2 + a^2 + 2a(n-c)p + (n-c)^2 p^2] \end{aligned}$$

MSE が p に依存しなければ、ミニマックスである。 p^2 の係数と p の係数がゼロという条件から $n = (n-c)^2$ および $n + 2a(n-c)p = 0$ となり、これから $c = n \pm \sqrt{n}$ および $a = \pm\sqrt{n}/2$ が得られる。 $a, c > 0$ だから $c = n + \sqrt{n}$, $a = \sqrt{n}/2$ とすると、ベイズ推定量は $\hat{p} = (x + \sqrt{n}/2)/(n + \sqrt{n})$ となり、これがミニマックス推定量で、 $MSE = \frac{n}{4(n+\sqrt{n})^2} = \frac{1}{4(\sqrt{n}+1)^2}$ である。グラフをみると $n = 100$ 程度より小さいときはミニマックスはそれほど悪くはない。

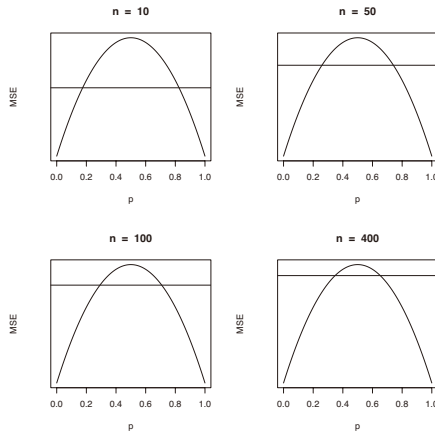


図 1. ミニマックスと x/n の MSE

謝 辞

本研究は JSPS 科研費 JP21H04403 の助成を受けた。

This work was supported by JSPS KAKENHI Grant Number JP21H04403.

参考文献

- [1] 美添泰人「ベイズの手法による統計分析 — 部分的なサーベイと今後の展望」, 東京大学出版会, 竹内 啓 (編)『計量経済学の新展開』第 6 章, 1983
- [2] 美添泰人「ベイズ統計学の原理と応用 (その 1)」『青山経済論集』第 74 巻 第 3 号, pp. 203–230, 2022
- [3] 美添泰人「ベイズ統計学の原理と応用 (その 2)」『青山経済論集』第 74 巻 第 4 号, pp. 119–169, 2023
- [4] Berger, J. O., *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer, 1985
- [5] Berger, J. O. and Wolpert, R., *The Likelihood Principle*, 2nd ed., IMS Lecture Notes– Monograph Ser. 9, Hayward, 1988
- [6] Bernardo, J. M., *Reference prior distribution for Bayesian inference* (with discussion), JRSS, B, 41, 113–147, 1979
- [7] A. Birnmbaum, “On the foundations of statistical-inference (with discussions),” JASA, 57, 269–326, 1962
- [8] Box. G. E. P. and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, 1973.
- [9] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw Hill, 1970
- [10] Gelfand, A. E., and Smith, A. F. M., “Sampling based approaches to calculating marginal densities,” JASA, 85, 398–409, 1990
- [11] Ghosh, J. K., M. Delampady and T. Smanta, *Bayesian Inference and Decision Theory*, Springer, 2006
- [12] Jeffreys, H., *Theory of Probability*, 3rd ed.(corrected), Clarendon Press, 1961, corrected impression 1967.
- [13] Lehmann, E. L., *Testing Statistical Hypotheses*, Wiley, 1959 (2nd ed. 1986)
- [14] Lindley, D. V., *An Introduction to Probability and Statistics from a Bayesian Viewpoint*, 1, 21, Cambridge Univ. Press, 1965
- [15] Lindley, D. V., *Bayesian Statistics, a review*, SIAM Monograph, 1972.

- [16] Lovric, M., (Editor-in-chief), *International Encyclopedia of Statistical Science*, Springer, 2010, 2nd ed. 2025
- [17] Pratt, J. W., Review of “Testing Statistical Hypotheses,” by E.L. Lehmann, JASA, 56, 163–168, 1961
- [18] Pratt, J. W., “Bayesian interpretation of standard inference statements,” JRSS, B, 27, 1965
- [19] Savage, L. J., *The Foundations of Statistics*, Wiley, 1954. Revised and enlarged edition, Dover Publications, Inc., 1972
- [20] Wald, A., *Statistical Decision Functions*, Wiley, 1950.