Japanese Psychological Review
2018, Vol. 61, No. 1, 101–115

ベイズファクターによる心理学的仮説・モデルの評価[1)]

岡　田　謙　介
東京大学

Evaluating psychological hypotheses and models using Bayes factors

Kensuke OKADA
The University of Tokyo

The Bayes factor has a basic and crucial role in Bayesian evaluation of psychological hypotheses and models. It forms a fundamental part of the advancement of psychological science. Its computation has been a major challenge, although recent advances in numerical estimation methods such as bridge sampling may allow the application of the Bayes factor to a wide range of practical research contexts. The objective of the current paper is to provide psychological scientists an introductory tutorial of the ideas and recent developments concerning the Bayes factor. Some running examples are presented and a few practical application methods are also discussed.

Key words: Bayes factor, Bayesian hypothesis testing, Savage-Dickey method, bridge sampling
キーワード：ベイズファクター，ベイズ的仮説検定，Savage-Dickey 法，ブリッジサンプリング

1. はじめに

ベイズ統計学は不確実性を確率によって表現し，観測したデータに基づいてその確率を更新する枠組みである。この更新はベイズの定理によって表現される。少数の原理原則を，幅広い問題に適用していくベイズ統計学の体系は，長年にわたり統計学の主流派ではなかったが，一部の熱意ある支持を集めてきた。近年，マルコフ連鎖モンテカルロ法の発展とそれを実装したソフトウェアの普及によってベイズ統計学の実用化が進み，心理学においてもベイズ統計学的な応用研究が隆盛の一途をたどっている。その一端は，心理学の主要な論文誌において，ここ 1，2 年だけでもベイズ統計学の特集号が相次いでいることにも見られる（Chow & Hoijtink, 2017；Etz & Vandekerckhove, 2018；Hoijtink & Chow, 2017；Mulder & Wagenmakers, 2016；van de Schoot, Schalken, & Olff, 2017）。これに対応して，わが国でも近年，心理学や関連分野への応用を主眼においたベイズ統計学の入門書やモデリングの書籍が相次いで出版されている（e.g., 伊庭，2018；Kruschke, 2014 前田・小杉（監訳），2017；Lee & Wagenmakers, 2013 井関（訳），2017；松浦，2016；奥村・牧山・瓜生，2018；豊田，2015, 2017）。本特集号においても，国里（2018），清水（2018），竹澤（2018），中村（2018）は心理学の諸分野におけるベイズ統計モデリングの有用性を論じており，また竹林（2018）はベイズ統計学に基づく柔軟な臨床試験デザインを紹介している。

量的な心理学研究は，リサーチ・クエスチョンに対する暫定的な答えとしての仮説を立て，モデルを構築し，これを得られたデータに基づいて吟味し検証することの積み重ねである（南風原，2011）。したがって，データに基づく仮説やモデルの評価は，研究を進める上で不可欠なステップとなる。ベイズ統計学の立場からこれを行うための，基本的であり，長い歴史を持ち，なおかつ近年の理論と応用両面における発展も著しい方法に，ベイズファクターを用いた仮説とモデルの評価がある。このアプローチは，本稿の以下で述べるようにベイズ統計学の自然な考え方に基づいており，従来型の仮説検定の欠点のうちいくつかを

1)　本研究は JSPS 科研費 17H04787 の助成を受けたものである。

克服し，そして大きな応用可能性を持つ。実際，*Journal of Mathematical Psychology* 誌の最近の特集号（Mulder & Wagenmakers, 2016）は，ベイズ統計学の中でも，とくにベイズファクターに対象を絞ったものであった。しかし，こうした近年の研究の進展を扱った，和文で入手可能なベイズファクターに関する総説は筆者の知る限り見当たらない。

そこで本稿では，この基本的で長い歴史を持ちながらも，現在新たな注目を集めているベイズファクターについて，心理学研究での応用を念頭におきながら，入門的なチュートリアルと総説を提供することを目的とする。

本稿の流れは以下の通りである。第2節では，比率と平均値についての推論問題を例にとり，ベイズファクターの定義と考え方について述べる。第3節では，ベイズファクターの特徴について，従来型の仮説検定法との比較も踏まえつつ述べる。第4節では，ベイズファクターを実際に計算するための方法として近年改めて注目されている，ブリッジサンプリングなどの数値的な方法について述べる。最後に，第5節では本論文を総括し，今後の展望について述べるとともに，実際の研究でデータからベイズファクターを算出するために利用可能なソフトウェアを紹介する。

2. ベイズファクターの基礎

2.1 比率のベイズ推論

本節では比率についての基本的なベイズ推論を導入し，次の2.2節でこの問題に対応する形でベイズファクターの考え方と解釈について述べる。

いま，関心のあるパラメータ（たとえば平均，分散，比率など）を θ で，データを y で表すことにする。本論文では，データとパラメータの一般表現としては上記を用いるが，具体的問題としてとりあげる比率や平均値の推論では，それぞれ文脈に合わせてデータとパラメータを表す変数を別途定義して用いるので注意してほしい。

ベイズ統計学におけるパラメータの推論で目的とするのは，データを得たもとでのパラメータ θ の確率分布である事後分布（posterior distribution）$p(\theta|y)$ を，ベイズの定理（Bayes' theorem）を用いてデータ y の情報から

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

と得ることである。ここで $p(y|\theta)$ はパラメータ θ を固定したときのデータ y についての確率分布である尤度（likelihood），$p(\theta)$ はデータ y が得られる前のパラメータ θ に関する不確実性を表す確率分布である事前分布（prior distribution），そして $p(y)$ は後に述べる周辺尤度（marginal likelihood）である。ここで，ベイズ統計学において完全な意味での統計モデルとは，データ y の生成メカニズムを表現する，(1) 式右辺分子の尤度 $p(y|\theta)$ と事前分布 $p(\theta)$ の組を指すことに注意する（e.g., Gelman et al., 2014；Little, 2006[2]）。

比率の推論のための具体的状況として，いま，問題に対して正答であれば y_i=1，誤答であれば y_i=0 という値をとる2値変数 y_i (i=1, ..., N) がデータであるとする。そしてこれは，

$$y_i \sim \text{Bernoulli}(\theta) \quad (2)$$

という，正答確率パラメータ θ を持つベルヌーイ分布から独立に得られた実現値と考えることにする。この N 個の観測値の情報をまとめて正答数 $k = \sum_{i=1}^{N} y_i$ として表すことにすると，その確率分布は

$$k \sim \text{Binomial}(N, \theta) \quad (3)$$

という二項分布で表すことができる。今回考えている比率の推論における (1) 式での尤度 $p(y|\theta)$ は，すべてのデータ $\{y_i\}$ (i=1, ..., N) に対する (2) 式，もしくは (3) 式の形で表されることになる。

ベイズ統計学ではパラメータに事前分布 $p(\theta)$ をおく。今回の状況において妥当な設定の1つは，データの情報を得る前には，正答確率 θ はその台でどの値も等しくとりうるというものであろう。すなわち，0から1までの一様分布

$$\theta \sim \text{Uniform}(0, 1) \quad (4)$$

を θ の事前分布に利用するということである。これは，

2) この論文は2005年のアメリカ統計学会における会長基調講演をまとめたものであり，ベイズ統計学と従来型の統計学の関係について一読の価値がある。

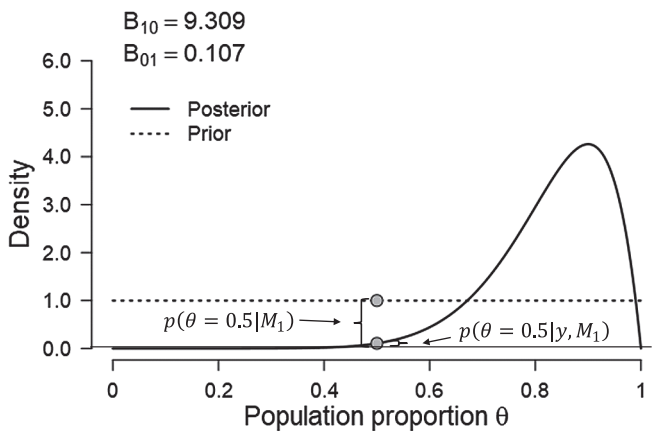

図1　比率の推論における事前分布（点線）と事後分布（実線），およびSavage-Dickey法で求められるベイズファクター $B_{10} = \dfrac{p(\theta=0.5|M_1)}{p(\theta=0.5|y, M_1)} = 9.309$

$$\theta \sim \mathrm{Beta}(1, 1) \tag{5}$$

というベータ分布としても表すことができる。

以上のように尤度と事前分布を与えたことにより，ベイズ統計モデルが1つ定められた。このモデルのもとで，データ $\{y_i\}$（もしくは，N と k）を観測してパラメータ θ についての推論を行うとき，θ の事後分布もまたベータ分布となる（たとえばKruschke, 2014 前田・小杉（監訳），2017を参照）。具体的には，この事後分布は

$$\theta \sim \mathrm{Beta}(k+1, n-k+1) \tag{6}$$

というベータ分布となる。

より具体的な例で考えるため，得られたデータが

$$\{y_i\} = \{1, 1, 1, 1, 1, 0, 1, 1, 1, 1\} \tag{7}$$

であったとする。つまり，N=10問の問題に対して正答が k=9回，誤答が $N-k$=1回であったということである。このとき，θ の事後分布は（6）式より

$$\theta \sim \mathrm{Beta}(10, 2) \tag{8}$$

となる。この事後分布を，事前分布とともに図1に示す。

2.2　比率のベイズ的仮説検定

前節で述べた比率の推論において，パラメータは正答確率 θ であった。このパラメータについて，従来型の方法を用いる場合には，

$$H_0: \theta = 0.5, \tag{9}$$
$$H_1: \theta \neq 0.5 \tag{10}$$

と，正答確率 θ がチャンスレベル（0.5）であるという帰無仮説 H_0 と，そうではないという対立仮説 H_1 を設定しての仮説検定がよく行われる。

これと同種の問題に，ベイズ統計学の枠組みで取り組むことを考えよう。ベイズ統計学ではパラメータ θ に事前分布 $p(\theta)$ をおくのであった。ここでパラメータ θ に関する（9）式は，

$$p(\theta = 0.5) = 1, \quad p(\theta \neq 0.5) = 0 \tag{11}$$

という一点分布の確率分布であると解釈できる。したがって，これを帰無仮説 H_0 に対応するモデルにおける事前分布として利用することができ，（2）式もしくは（3）式で表される尤度と合わせて，帰無仮説に対応する統計モデルが1つ定まる。これを M_0 と表すことにする。

一方，対立仮説 H_1 の（10）式を満たす θ の確率分布は，一意には定まらない。そこで具体的な事前分布の形を決める必要がある。応用文脈上，もっとも自然な設定の1つは，データを得る前の θ に対して一様分布をおく（4）式の設定であろう（前述の通り（5）式もまったく同じ分布を表す）。そこでこれを事前分布とし，（2）式もしくは（3）式で表される尤度を利用することにすると，対立

仮説 H_1 に対応する統計モデル M_1 も 1 つ定まる。これはもちろん，2.1 節で推論に用いたモデルと同じである。(4) 式の一様分布は一見，θ が 0.5 以外の値であるという (10) 式とは矛盾するように見えるかもしれないが，連続型確率分布では確率変数が任意のある 1 点の値をとる確率は 0 であり，したがって (4) 式の設定のもとで θ が 0.5 の値をとる確率は（ほかのどの値をとる確率とも同様に）0 である。このように，モデル M_1 のもとで (10) 式の条件は満たされている。

ベイズ的仮説検定（Bayesian hypothesis testing; e.g., Wagenmakers et al., 2010）とは，ここで導入したように，2 つの仮説を表現する統計モデル M_0，M_1 を定め，ベイズファクターを用いて両者を比較する枠組みのことをいう。すなわち，今回の例では

$$M_0: \begin{cases} y_i \sim \text{Bernoulli}(\theta) \\ \theta = 0.5 \end{cases} \tag{12}$$

と

$$M_1: \begin{cases} y_i \sim \text{Bernoulli}(\theta) \\ \theta \sim \text{Uniform}(0, 1) \end{cases} \tag{13}$$

という 2 つのモデル間で，データに基づいてモデル比較を行うこととなる。この 2 つのモデル間で異なるのはパラメータ θ についての設定なので，ベイズ的仮説検定でも (9)，(10) 式のように両モデルについて異なるパラメータだけについての表記がされることもある。しかし，ベイズファクターによって実際に比較されるのは (12)，(13) 式のような尤度と事前分布の組によって定まる，2 つのベイズ統計モデルであることに注意してほしい。ベイズ的仮説検定では，仮説は尤度と事前分布からなる統計モデルによって表現されるのである。

あらゆるベイズ推論は，特定の統計モデルのもとで行われる推論である。そこで，たとえば (13) 式のモデル M_1 のもとでベイズの定理による通常のベイズ推論（(1) 式）を行うとき，このモデル M_1 の設定のもとでの（すなわち，M_1 で条件付けられた）推論であることを明示的に条件付ける形で示して

$$p(\theta|y, M_1) = \frac{p(y|\theta, M_1)p(\theta|M_1)}{p(y|M_1)} \tag{14}$$

と表現することもできる。

ここで，分母の $p(y|M_1)$ は周辺尤度もしくはエビデンス（evidence）と呼ばれる量である。文脈によっては，敢えてモデルで条件付ける表記をせずに，(1) 式のように周辺尤度を $p(y)$ と書くことも多い。パラメータ空間の全体を Θ で表すとき

$$p(y|M_1) = \int_\Theta p(y|\theta, M_1)p(\theta|M_1)d\theta \tag{15}$$

が成り立つことから，周辺尤度はこのパラメータ θ のとりうるすべての値について考慮したときの，手元のデータ y に対するモデル M_1 の持つ平均的な予測力を表すと解釈できる。すなわち，周辺尤度の大きいモデルほど，相対的にデータを説明する上でふさわしいと解釈できる。もちろん，(14)，(15) 式は M_1 を M_0 に置き換えても同様に成り立つ。そこで，モデル M_1 と M_0 の 2 つのモデルを比較するにあたり，周辺尤度の比

$$B_{10} = \frac{p(y|M_1)}{p(y|M_0)} \tag{16}$$

をとり，これが 1 より大きければ M_1 が，また 1 より小さければ M_0 が，データ y を説明する上でよりふさわしいと支持されたのだと考えることができる。また，その 1 より大きい／小さい程度が，モデルがデータを説明・予測する上でのふさわしさの程度であると解釈することができる。このような解釈ができる (16) 式の B_{10} が，H. Jeffreys (1935, 1961) が導入したベイズファクター（Bayes factor）である[3]。日本語ではベイズ比，ベイズ因子，ベイズ因数と訳されることもある（南風原，2014；繁桝，1995）。ベイズファクターは 2 つのベイズ統計モデルの周辺尤度の比であり，これを用いて，モデルがデータをよく説明・予測する度合いの相対的比較を行うことができる。その値の具体的な解釈の目安としては，3.2 未満であれば M_0 と比べて M_1 を支持する傾向があるとはいえてもあまり実質的意味があるとは考えにくいが，3.2 よりも大きければその支持には実質的な意味があり，また 10 より大きければ強い支持を与える，としたH. Jeffreys（1961）による基準や，それを修正した Kass and Raftery（1995）による基準がよく

3)　Alan Turing も第 2 次対戦中のドイツ軍の暗号を解読するための極秘研究の中で，ベイズファクターの概念を提案していたことが知られている（Pericchi, 2005）。

知られている。もちろんベイズファクターの値は個々の応用例に則して解釈されるべきであるが，これらは1つの解釈の目安・ものさしを与えるものである（繁桝，1995）。

ところで，ベイズ統計学の基本的な考え方は，不確実性を確率で表現し，データの情報に基づいてこの確率を更新することであった。そこで，2つのモデル M_0 と M_1 の間で不確実性があるのであれば，パラメータについての推論の場合とまったく同様に，ベイズの定理を使って，データを得たもとでのモデルについての確率の更新を

$$p(M_0|y) = \frac{p(y|M_0)p(M_0)}{p(y)}, \tag{17}$$

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)} \tag{18}$$

と行うことができる。この $p(M_0|y)$ や $p(M_1|y)$ を事後モデル確率（posterior model probability）という。そして，この2つの事後モデル確率の比をとると，分母が打ち消し合って，

$$\frac{p(M_1|y)}{p(M_0|y)} = \frac{p(y|M_1)p(M_1)}{p(y|M_0)p(M_0)} = B_{10}\frac{p(M_1)}{p(M_0)} \tag{19}$$

という関係式を得ることができる。確率の比をオッズと呼ぶので，この式の左辺 $p(M_1|y)/p(M_0|y)$ は，2つのモデル間の，データの情報によって更新されたオッズを表し，事後モデルオッズ（posterior model odds）という。同様に，右辺の $p(M_1)/p(M_0)$ の項は，データを得る前における2つのモデル間のオッズを表し，事前モデルオッズ（prior model odds）という。そして，（19）式で事前モデルオッズを事後モデルオッズへと変換する際にかかるのが，（16）式のベイズファクター B_{10} である。すなわち，ベイズファクター B_{10} は，事前モデルオッズから事後モデルオッズへの変化の度合いを，比の形で表現した量となっている。このことから，ベイズファクターは，モデルのオッズ比を基準としてデータがもたらす「証拠の重さ」を表す量とも解釈することができる（Good, 1985）。ベイズファクター B_{10} が1より大きいとき，データの情報はモデル M_1 の事後モデル確率を事前よりも大きくする方向にはたらき，1より小さいときはその逆となる。このように，ベイズファクターは周辺尤度の比であることに加えて，事後オッズと事前オッズの比としても解釈することができる。

ここまで，ベイズファクターの2つの代表的な解釈について述べてきた。ベイズファクターはベイズ統計学の基本的な体系だけから直接的に導かれる量であり，少数の原理原則を幅広い問題に適用していくベイズ統計学の考え方と整合する概念であるといえる。前提の少なさに由来する汎用性は，ベイズファクターの大きな魅力の1つである。

2.3 Savage-Dickey 法

定義にしたがってベイズファクターを計算するには，比較する両モデルそれぞれについて（15）式の周辺尤度を求めることになる。前節で扱った比率に関する2つのモデルのような単純なモデルの場合であれば，この積分計算は難なく可能である。しかし，より複雑なモデルのもとでは，パラメータ θ が多次元になり，（15）式に対応する（多重）積分が解析的に求められない場合が多くなる。

Gelfand and Smith（1990）以降，事後分布から多数の乱数をサンプリングし，この乱数を用いてベイズ推論を行うマルコフ連鎖モンテカルロ（Markov chain Monte Carlo, MCMC）法の発展が，ベイズ統計学を実用的な方法論へと大きく変えた。MCMC 法を用いれば，事後分布自体が解析的に求められなくても事後分布から発生させた乱数のヒストグラムを利用することができ，また事後分布の平均（事後平均）が解析的に求められなくてもこの乱数の平均値を利用することができる。MCMC 法によって得られるのは，もちろん事後分布やその要約統計量自体ではなく，多数の乱数に基づくその近似値である。しかし現代の計算機を用いれば，事後分布から多数の乱数をわずかな時間でとりだすことができる。そして，こうした乱数による近似は多くの場合，応用上十分な精度を持つ。したがって，乱数発生に基づく事後分布やその要約統計量の推定は，その汎用性と精度の点で実用的かつ強力な方法といえる。一方，4節で改めて論じるが，単純な乱数発生に基づく周辺尤度の推定は，確率分布の裾の領域の評価精度に問題があり，残念ながら実用的な方法とはならない。

ここで，2.2節で扱ったモデル M_0 と M_1 の間に

は，モデル M_1 がパラメータ θ を推定するのに対して，M_0 では θ=0.5 と制約するという関係にあった。このように，一方のモデル M_1 のパラメータ空間に制約を加えることで他方のモデル M_0 が得られるとき，モデル M_0 は M_1 にネストしている（nested）という。そして，等号制約によって一方が他方にネストした関係にある2つのモデル間のベイズファクターは，非常に簡単な計算によって求めることが可能である。すなわち，図1に例示するように，M_0 で制約するパラメータの値（今回でいえば θ=0.5）における M_1 の事前確率密度と事後確率密度の比

$$B_{10} = \frac{p(\theta=0.5|\, M_1)}{p(\theta=0.5|y, M_1)} \tag{20}$$

が，モデル M_0 に対してモデル M_1 を支持するベイズファクター B_{10} となるのである。この（20）式の右辺を Savage-Dickey 密度比（Savage-Dickey density ratio）といい，これを用いてベイズファクターを求めることを Savage-Dickey 法という（Verdinelli & Wasserman, 1995；Wagenmakers et al., 2010）。この結果は Dickey and Lientz（1970），Dickey（1971）によるものであり，彼らは Savage の未発表原稿を初出としている。

具体的に考えると，2.2 節の状況において，（7）式のデータのもとでモデル M_0 と M_1 とを比較するベイズファクターは，図1に示した $p(\theta=0.5|M_1)$ と $p(\theta=0.5|y, M_1)$ の比によって，B_{10}=9.309 と求めることができる。

このように，等号制約によってネストした2つのモデル間の比較は，単に事後分布と事前分布の確率密度の比を求める単純な問題に帰着するのである。そして，この事後分布と事前分布の確率密度値の算出は，MCMC 法のような乱数発生に基づく方法とカーネル密度推定によって行うことができる。このように汎用性が高いことから，実際の心理学研究におけるモデル比較でも，Savage-Dickey 法を用いてベイズファクターが算出される例は多い（e.g., Okada, Vandekerckhove, & Lee, 2018；Wagenmakers, Verhagen, & Ly, 2016）。

2.4　平均値のベイズ的仮説検定

比率の比較と並んで心理学の応用上よく取り上げられる問題に，平均値の比較がある。本節ではこれを考える。

尤度として，心理学の典型的な応用で用いられるように，観測データ $\{y_i\}$ が平均 μ，分散 σ^2 の正規分布から得られた独立な観測値であるという設定

$$y_i \sim \text{Normal}(\mu, \sigma^2) \tag{21}$$

を考えることにする。ここで，従来型の方法を用いる場合には，

$$H_0: \mu = 0, \tag{22}$$
$$H_1: \mu \neq 0 \tag{23}$$

と，平均 μ が0であるという帰無仮説 H_0 と，そうではないという対立仮説 H_1 を設定しての検定がよく行われる。ベイズ統計学の枠組みにおいても，帰無仮説を表現するモデル M_0 では，比率の場合と同様に μ=0 の一点分布を事前分布に利用できる。一方，対立仮説側のモデル M_1 では μ を推論することになる。ここで，ベイズファクターを用いてモデル比較を行う場合には，モデル M_1 のもとで μ について，そのとりうる全範囲に対する一様事前分布を設定することは妥当ではない。なぜならば，平均 μ は理論上の台が $[-\infty, \infty]$ であるため，一様分布は積分が1にならず，確率分布の定義を満たさない非正則な（improper）分布になり，そのために μ=0 の点における確率密度の値（すなわち図2における高さ）を一意に決めることができないからである。したがって，M_1 のもとでの μ についての事前分布は，何らかの理論に基づいた，非正則でない分布を設定する必要がある。

ここでは Morey et al.（2011），Rouder et al.（2009）にならい，心理学の応用において現在もっともよく利用される設定の1つである Jeffreys-Zellner-Siow（JZS）の事前分布を考える。この方法では，まず事前分布が y の単位に依存しないようにするために，

$$\delta = \frac{\mu}{\sigma} \tag{24}$$

という新たな（決定的な）パラメータを考える。ここで μ=0 のとき δ=0 であることに注意する。この δ は，効果量の考え方と同様に，データ y のしたがう分布の標準偏差 σ を単位として平均 μ を

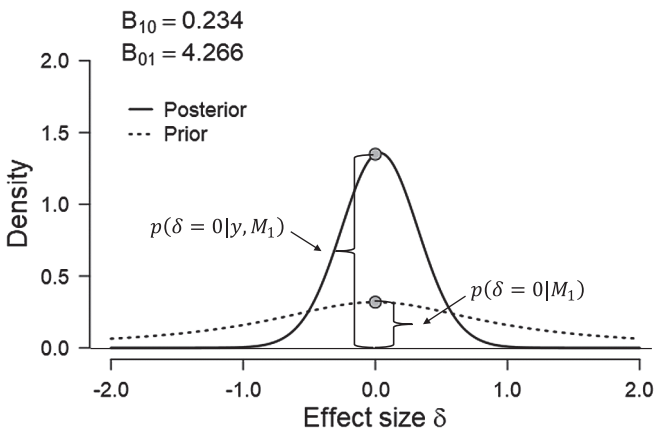

図2　平均値の推論における事前分布（点線）と事後分布（実線），およびSavage-Dickey法で求められるベイズファクター $B_{10}=\frac{p(\delta=0|M_1)}{p(\delta=0|y,M_1)}=0.234$

表現しなおした量になっている。そして，この効果量に相当するパラメータ δ に対して

$$\delta \sim \text{Cauchy}(1) \tag{25}$$

と標準コーシー事前分布をおく。この設定は，H. Jeffreys（1961）が，導出されるベイズファクターが統計学的に望ましい性質（デシデラータ desiderata と呼ばれる）を満たすような事前分布の中で，もっともシンプルな分布として採用したものである。

もう1つのパラメータ σ^2 は比較する2つのモデル間で共通である。こうしたパラメータについては，その事前分布の設定が，結果として得られるベイズファクターにあまり影響しないことが知られている。ここでは，やはりH. Jeffreys（1961）にならい，いわゆるJeffreysの事前分布と呼ばれる

$$p(\sigma^2) \propto 1/\sigma^2 \tag{26}$$

を利用することにする。

上記のように，δ にコーシー分布を，σ^2 にJeffreysの事前分布を設定するアプローチを，Bayarri and Garcia-Donato（2014）はH. Jeffreys（1961）とZellner and Siow（1980）に帰してJeffreys-Zellner-Siow（JZS）の事前分布と呼んだ。このJZS事前分布のもとでベイズファクターは，積分を含むが計算可能な形で導出できる（Rouder et al., 2009の(1)式を参照）。

具体的な例として，

$$\{y_i\} = \{-1.7, 1.6, 0.3, -0.5, 0.3, 0.2, -0.2, -0.9, 0.8, 0.5\} \tag{27}$$

というデータが得られた場合の δ の事前分布と事後分布，Savage-Dickey密度比を図2に示す。Savage-Dickey法より，ベイズファクターは B_{10}=0.234（=1/4.266）となり，データは帰無仮説に対応するモデル M_0 に対して支持を与えていることがわかる。

3. ベイズファクターの特長と留意点

ここでは，前節で導入したベイズファクターの性質について，とくに心理学での応用に関係が深いと考えられる点をまとめる。3.1節ではその特長について，3.2節では留意すべき点について述べる。

3.1　ベイズファクターの特長

(1) 自然で基本的な量である

これまでに述べてきたように，ベイズファクターは，不確実性を確率によって表現し，それをデータに基づいて更新するというベイズ統計学の枠組みの中で自然に得られる，基本的かつ本質的な量である。Good（1995）は*Nature*誌の論考を，「ベイズファクターという易しい概念は裁判の基礎である。また医学診断の基礎であり，科学的思考の基礎でもある。大学入学前に教えるべき

だ！」（拙訳）と締めている。

(2) 帰無仮説を支持できる

これまで主流であった，p 値を用いた従来型の仮説検定の問題点の 1 つとして，仮説に関する非対称性が挙げられる。p 値は，帰無仮説が真であることを前提として導かれる量である。したがって，p 値が小さい場合には，前提となった H_0 を，第 1 種の誤りの確率を認めた上で棄却することができる。しかし，p 値は帰無仮説 H_0 に反する証拠の大きさを表す量ではないため，p 値が大きいこと，H_0 が棄却できないことは，対立仮説 H_1 を積極的に支持する証拠とはいえない（Wilkinson & Task Force on Statistical Inference, 1999）。このような，仮説に関する非対称性に由来する問題は，心理学における統計改革の中で p 値を用いた従来型の仮説検定が批判される大きな理由の 1 つであった（大久保・岡田，2012）。

一方で，ベイズファクターはモデル M_1 と M_0 の周辺尤度の比であり，帰無仮説側のモデル M_0 に特別な地位を与えていない。M_1 を M_0 と比較するベイズファクター B_{10} は，定義より，単に

$$B_{10} = \frac{1}{B_{01}} \tag{28}$$

と，M_0 を M_1 と比較するベイズファクター B_{01} の逆数になる。比較される 2 つのモデルは平等であり，ベイズファクターは帰無仮説に対応するモデル M_0 を支持する証拠も，それに反する証拠もまったく同じように与えることができる。実際に，ベイズファクターを用いることで帰無仮説側を支持する心理学研究は多数出版されている（e.g., Dienes, Coulton, & Heather, 2018；Wetzels et al., 2009）。

(3) 逐次的更新ができる

従来型の仮説検定は，多くの場合，サンプルサイズ N を定数に固定して構成されている。この前提が満たされていない場合には，明示的にそのことを考慮した，通常使われるのとは異なる仮説検定法を利用しなければいけない。たとえば，100 人分のデータを収集し，仮に有意でなければもう 100 人を追加する研究デザインから，200 人分のデータが収集されたとする。このとき，このデータを通常の N=200 のデータと考えて仮説検定を行うと，第 1 種の誤りの確率が本来の設定よりも上昇してしまう。これは，データ収集デザイン（サンプリングデザイン）と，統計分析法（検定法）の前提とが異なるためである。したがって，従来型の枠組みにおける仮説検定では，事前に検定すると決めた人数分のデータを集めてから，はじめて検定を行う必要がある（この人数の決め方，すなわちサンプルサイズ設計については村井・橋本，2017, 2018 に詳しい）。

従来型の枠組みの中でも，Wald（1945）以降，人数を足しながら検定を繰り返していくデータ収集デザインに対応した逐次検定（sequential testing）法は開発されている。しかし，医学分野では多く利用されているものの，心理学では普及をみせていない。また，こうした方法を用いる場合でも従来型のアプローチをとる限り，事前にサンプリングデザインを決め，その計画どおりにデータを収集する必要があることに変わりはない。実際の心理学研究では，本来逐次検定が必要な状況において厳密な第一種の誤りの確率の統制をすることなく検定が使われている場合も多いと考えられる。このことは，false-positive psychology と言われるような，再現性の問題を招く原因の 1 つとして指摘されている（三浦，2017；大久保，2016；Simmons, Nelson, & Simonsohn, 2011）。

一方で，ベイズ統計学の枠組みでは，ベイズファクターが定めた閾値に達するまでデータを順次追加していき，達した時点でデータ収集をやめることが，通常の方法を用いて問題なく可能である（Dienes, 2008, 2011；Kass & Raftery, 1995；Rouder, 2014；Schönbrodt et al., 2017）。この違いは，従来型の枠組みにおけるパラメータの推論が，観測される可能性があった（が観測されなかった）データに依存するのに対して，ベイズ統計学では実際に観測されたデータだけで条件付けた推論を行えばよいことに由来する。この 2 つの枠組みの違いについての具体例としては，Wagenmakers（2007）の Online Appendix が参照できる。

(4) 同程度の予測力ならば単純なモデルを選ぶ

周辺尤度は，データを予測する度合いがもし 2 つのモデル間で同程度であるならば，複雑なモデルよりも単純なモデルにおいて大きくなる性質を持つ。多数のパラメータを持つ複雑なモデルの周辺尤度が，単純なモデルのそれよりも大きく

なるのは，前者のモデルのデータに対する予測力が後者よりも十分に高いときに限られる。この性質は W. H. Jeffreys and Berger（1992）や Lee and Wagenmakers（2013 井関（訳），2017）に，具体的な数値例とともに示されている。したがって，周辺尤度の比として得られるベイズファクターも，データに対して過度に複雑なモデルよりも，シンプルにデータを説明できるモデルを支持する傾向を持つ。Smith and Spiegelhalter（1980）は，ベイズファクターの持つこの性質を，自動的なオッカムの剃刀（automatic Ockham's razor）と呼んでいる。

(5) 望ましい理論的性質を持つ

モデル選択の一致性（consistency）とは，比較するモデルの中に真のモデルが含まれていれば，十分大きなサンプルサイズのもとで，その真のモデルが選択される確率が十分に 1 に近くなるという性質である。たとえば 2.4 節で述べた Jeffreys-Zellner-Siow の事前分布を設定したもとで，ベイズファクターには一致性がある（Bayarri et al., 2012；Jeffreys, 1961）。すなわち，サンプルサイズが十分大きくなれば，ベイズファクター B_{10} はモデル M_1 が真のとき無限大に，モデル M_0 が真のとき 0 に収束する。これをはじめとして，典型的な設定のもとでのベイズファクターは統計学的にみて望ましい，複数の性質を満たしている（Berger, 2006b；Kass & Raftery, 1995）。

3.2 ベイズファクターの留意点

(1) 相対比較である

ベイズファクターは周辺尤度からみた，2 つのモデルの観測データに対する相対的な予測力の比であり，モデルの絶対的な適合の度合いを表すわけではない。したがって，たとえばベイズファクターが 100 であったとき，すなわち一方のモデルが他方のモデルに比べて周辺尤度の意味で 100 倍支持されたときでも，そもそもどちらのモデルもデータ生成メカニズムの表現としてはまったく不十分ということはありうる。すなわち，モデルの絶対的なあてはまりのよさについては，ベイズファクター以外の観点からも検討される必要がある。

ベイズ統計モデルの絶対的なあてはまりを評価するときの，基本的な方法は事後予測チェック（posterior predictive checking; Gelman, Meng, & Stern, 1996）である。これは，適切なモデルであれば，将来観測されるであろうデータについての確率分布である事後予測分布と，実際に観測されたデータの間には類似性が見られるであろうという考えに基づき，両者を比較するものである。視覚的に事後予測チェックが行われることもあれば，事後予測 p 値（posterior predictive p value; Meng, 1994）などを用いて定量的な検討が行われることもある。MCMC 法を用いた推定では，事後分布からだけでなく事後予測分布からの乱数を発生させることも容易に可能であり，これによって事後予測チェックを行うことができる。

(2) 事前分布の設定も影響する

ベイズファクターを用いて比較されるのは，事前分布と尤度からなる統計モデルである。2.4 節で述べたように，比較されるうちの一方がパラメータを特定の値に固定する点帰無仮説に対応するモデル M_0 である場合には，対立仮説に対応するモデル M_1 において，そのパラメータの事前分布をどのように設定するかが重要となる。それは，Savage-Dickey 法の考え方から理解できるように，この事前分布の設定が，得られるベイズファクターの値に大きく影響しうるからである。たとえば 2.4 節で考えた平均値に関するモデルの比較において，モデル M_1 で平均 μ の事前分布の分散を大きくすると，ベイズファクターはモデル M_1 よりも M_0 を支持するようになる。ただし，2 つのモデル間で共通なパラメータ（4.2 節の例でいえば σ^2）に対する事前分布の設定は，ベイズファクターに対して通常あまり大きな影響を与えないことが知られている。

この，事前分布の設定に対してベイズファクターが敏感でありうるという性質は，ベイズファクターの欠点ではなく，むしろ特長の 1 つだと考えることができる（Lee & Wagenmakers, 2013 井関（訳），2017）。それは，事前分布は尤度と並んでベイズ統計モデルの重要な一部であり，過度に分散の大きな事前分布は，たとえばデータが身長であるときに，165 cm も 16500 cm も −4000 cm も同程度の確率で得られうるといったような，現実的にみて妥当ではないモデルの設定をしていることになるからである。また，応用上利用可能な情報を事前分布として取り出す（確率の抽出 elicitation と呼ばれる）ためのさまざまな方法も提案されている（Albert et al., 2012；Jenkinson, 2005；Jones

& Johnson, 2014；Savage, 1971）

一方で，モデルのよい事前分布を構成するために利用できる情報が，いつも入手可能だとは限らない。そこで応用上，対立仮説を表現するモデル M_1 で標準的に利用できる，研究者の間で合意がとれた事前分布があると好都合である。このような目的から，客観ベイズ（objective Bayes）の研究分野では，汎用性のある事前分布の設定法が盛んに研究されてきている（Berger, 2006a）。このときの基本的な考え方は，事前分布が満たすべき性質（デシデラータ desiderata）をリストアップし，それを満たすような中でシンプルな事前分布を数学的に導出するというものである。このようにして導かれた事前分布を，既定事前分布（default prior）という（Berger & Pericchi, 1996；Ly, Verhagen, & Wagenmakers, 2016）。たとえば 2.4 節で導入した Jeffreys-Zellner-Siow の事前分布も，こうした客観ベイズ的考え方に基づいて導かれた既定事前分布の一つである。先行研究では典型的な応用場面で利用できる各種の既定事前分布が提案されている。その一部は，5 章で紹介するソフトウェアでも利用できるようになっている。

(3) 計算が必ずしも容易でないことがある

ベイズ統計モデリングの応用研究の増加と比して，ベイズファクターの応用がこれまで大きくは進んでこなかった理由の 1 つに，計算上の問題がある。

ベイズファクターは，3.1 節で述べたような利点を持つ。また 2.2，2.4 節の状況をはじめとした比率や平均値などの定型的な問題や，Savage-Dickey 法が使える問題については，比較的容易に算出が可能である。しかし，心理学や社会科学の応用で多く利用されるような，現実的な複雑さを持ったモデル間の比較においては，周辺尤度を，そしてベイズファクターを実用上十分な精度で求めるための汎用的方法がなかった。このことは，WAIC（Watanabe, 2010）や LOO（Vehtari, Gelman, & Gabry, 2017）といった，MCMC 法で得られた乱数値から汎用的に計算できる，将来のデータの予測や汎化性能の観点からモデルを評価する指標が近年よく利用される 1 つの理由になってきた。

この問題に対して，近年，ブリッジサンプリングを用いたベイズファクターの算出が，その汎用性から注目されている。次の 4 節では，これを含め数値的にベイズファクターを求める方法について，Gronau, Sarafoglou et al.（2017）に基づいて述べる。

4. ブリッジサンプリングと数値的方法

4.1 ナイーブ・モンテカルロ法

ナイーブ・モンテカルロ（naive Monte Carlo）法は，周辺尤度を与える（15）式の積分を，そのまま乱数に基づく数値積分に置き換える方法である。簡単のためモデルで条件付ける表記を省くと，ナイーブ・モンテカルロ法による周辺尤度推定の手順は，模式的に

$$\tilde{\theta}_i \sim p(\theta), \tag{29}$$

$$\hat{p}_1(y) = \frac{1}{R}\sum_{i=1}^{R} p(y|\tilde{\theta}_i) \tag{30}$$

と書くことができる。すなわち，(29) 式のように事前分布 $p(\theta)$ から θ の乱数 $\tilde{\theta}_i$ を発生させて，その値で尤度関数 $p(y|\theta)$ を評価することを多数回（R 回）繰り返し，(30) 式のようにその平均をとることで周辺尤度 $p(y)$ を推定する。

この方法は簡便であり，MCMC 法と同様に理論上もうまく機能しそうに思われる。実際，周辺尤度のナイーブ・モンテカルロ推定量は，事前分布と事後分布の形状が近く，重なりが大きいときには適切に機能する。しかし，残念ながら，事後分布が事前分布よりも軽い裾と鋭い峰を持つ場合には，(30) 式で周辺尤度を推定する際の標準誤差が大きくなってしまうことが知られている。そして，多くの応用場面では，データの情報によって更新された事後分布は，事前分布よりも軽い裾と鋭い峰を持つ。こうした場合，(29) 式で事前分布から乱数発生された $\tilde{\theta}_i$ の大半が尤度の極めてゼロに近い領域に入ることになる。その結果，(30) 式の和を評価する際に大半の尤度 $p(y|\tilde{\theta}_i)$ はほぼ 0 になり，まれに大きな値となって，その平均が安定しなくなってしまうのである。

4.2 重点サンプリング法

重点サンプリング（importance sampling）は，上述のナイーブ・モンテカルロ法の問題を解決するため，事前分布 $p(\theta)$ ではなく重点分布 $g(\theta)$ か

らの乱数発生値を利用する。この重点分布は，(1) 事後分布より裾が重く (2) 事後分布と形状が類似しており (3) 事後分布と同じ台を持ち (4) 評価が容易な分布であるように選ぶ（Neal, 2001；Vandekerckhove, Matzke, & Wagenmakers, 2015）。尤度の大きな部分からは頻繁に乱数値のサンプリングが行われ，逆に尤度の小さな部分からはまれにしかサンプリングがされなくなるようにすることで，周辺尤度の推定が不安定になる問題が生じにくくなるのである。ナイーブ・モンテカルロ法の場合と同様の模式的表記を用いると，重点サンプリングによる周辺尤度推定の手順は

$$\tilde{\theta}_i \sim g(\theta), \tag{31}$$

$$\hat{p}_2(y) = \frac{1}{R}\sum_{i=1}^{R}\left(p(y|\tilde{\theta}_i)\frac{p(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}\right) \tag{32}$$

と書くことができる。すなわち，直接事前分布 $p(\theta)$ から乱数発生をするかわりに，まず重点分布 $g(\theta)$ からのサンプリングを行い，これを使って尤度 $p(y|\theta)$ を評価した上で，事前分布と重点分布の比をかけて周辺尤度に対応する量に戻す。これを多数集めて平均をとることによって，重点サンプリングによる周辺尤度の推定値を得ることができるのである。

重点サンプリングは適切な重点分布 $g(\theta)$ が得られれば，ナイーブ・モンテカルロ法よりもはるかに効率よく，精度よく周辺尤度を推定できる。一方で，重点分布は前述の条件を満たすように，問題に合わせて選ぶ必要がある。とくに，事後分布と類似していてかつ裾が重くなるように重点分布を選ぶことは必ずしも容易でなく，汎用的な方法としてアルゴリズムを実装することには困難があった。

4.3 一般化調和平均サンプリング法

一般化調和平均サンプリング（generalized harmonic mean sampling）では，重点分布ではなく事後分布 $p(\theta|y)$ からサンプリングした乱数値を用いて周辺尤度を推定する。その手順は

$$\theta_j^* \sim p(\theta|y)\,, \tag{33}$$

$$\hat{p}_3(y) = \left(\frac{1}{R}\sum_{j=1}^{R}\left(\frac{1}{p(y|\theta_j^*)}\frac{g(\theta_j^*)}{p(\theta_j^*)}\right)\right)^{-1} \tag{34}$$

と書くことができる。一般化調和平均サンプリングでも重点分布 $g(\theta)$ は利用されるが，これは，(1) 事後分布より裾が軽く (2) 事後分布と形状が類似しており (3) 事後分布と同じ台を持ち (4) 評価が容易な分布であるように選ぶ（Newton & Raftery, 1994）。すなわち (1) の裾の軽重だけが逆であり，(2)～(4) は重点サンプリングと同じである。(34) 式でとられている総和は，(32) 式の逆数の形になっており，重点サンプリングと対比される方法であることが見てとれる。

4.4 ブリッジサンプリング法

重点サンプリングと一般化調和平均サンプリングは，いずれもナイーブ・モンテカルロ法の精度の問題を改善した数値的方法であるが，分布の裾に関して強い条件が必要であった。ブリッジサンプリング（bridge sampling；Meng & Wong, 1996）は，この裾に関する制約を緩めた方法である。

ブリッジサンプリングによる周辺尤度推定の手順は

$$\tilde{\theta}_i \sim g(\theta), \tag{35}$$

$$\theta_j^* \sim p(\theta|y), \tag{36}$$

$$\hat{p}_4(y) = \frac{\frac{1}{R_2}\sum_{i=1}^{R_2} p(y|\tilde{\theta}_i)p(\tilde{\theta}_i)h(\tilde{\theta}_i)}{\frac{1}{R_1}\sum_{j=1}^{R_1} h(\theta_j^*)g(\theta_j^*)} \tag{37}$$

と書くことができる。ブリッジサンプリングでは，提案分布 $g(\theta)$ からの乱数 $\tilde{\theta}_i$ と，事後分布 $p(\theta|y)$ からの乱数 θ_j^* の両者を利用する。この提案分布 $g(\theta)$ は，事後分布と類似していて重なりが大きいように選ばれる分布であり，その点で重点分布と似ている。実際上よく提案分布として利用されるのは，平均と分散を事後分布と揃えた正規分布である（Overstall & Forster, 2010）。また，高次元などより困難な状況に対しては，さらに洗練された方法も提案されている（Fruhwirth-Schnatter, 2004；Wang & Meng, 2016）。

また，$h(\theta)$ はブリッジ関数（bridge function）と呼ばれる関数である。典型的には，Meng and Wong (1996) の提案した，事前分布・提案分布・事後分布・周辺尤度の 4 つの情報を基に平均二乗誤差を最小化する最適ブリッジ関数（optimal bridge func-

tion）が利用される。この最適ブリッジ関数は推定すべき周辺尤度にも依存した形をしているため，実際の応用上は周辺尤度を逐次的に更新する反復型のアルゴリズムによって推定が行われる。

このように，ブリッジサンプリングは以前の方法と比べてアルゴリズムは複雑になるが，分布の裾に対する厳しい条件を外すことができるという大きな利点がある。また，MCMC 法によって得られたパラメータの乱数値から，ブリッジサンプリングを用いて周辺尤度の推定を行う汎用性の高い R パッケージ bridgesampling が開発され（Gronau, Singmann, & Wagenmakers, 2017），注目を集めている。ベイズファクターの推定に利用できる数値的方法の，歴史的経緯や具体例を含むさらなる解説としては Gelman and Meng（1998），Gronau, Sarafoglou, et al.（2017）が参照できる。

5. まとめと実用のために

本稿では，ベイズ統計学に基づいた仮説やモデルの評価のための，基本的で，しかし重要な量であるベイズファクターについて，その考え方を簡単な例に基づいて紹介し，特長と留意点をまとめ，そして応用への普及が期待される数値的な推定法を紹介した。

近年のベイズ統計学の隆盛は，その実用性の向上と密接な関係がある。ベイズ統計学が自然な考え方に基づき，大きな可能性を持つことは，研究者の間では従来から繰り返し指摘されてきた。しかし，1980 年代以前は実際的な問題にベイズ推定を適用することが必ずしも容易でなかった。この状況が一変したのは，Gelfand and Smith（1990）以降の MCMC 法の発展と，それを実装した汎用的なソフトウェアの普及（Lunn et al., 2009）によるものである。現代では，ベイズ統計学は実用的な体系として研究者や実務家の間で広く認識されるようになった。

これと同様に，ベイズファクターも，最近までは単純な問題や Savage-Dickey 法が利用できる状況を除いて汎用的にモデル評価のために利用することは困難であった。しかし，本論文で概観してきたように，近年理論的および数値的な方法についての研究が進展してきており，その実用性が高まっている。

実際，手元のデータに対してベイズファクターを用いて典型的なモデル間の比較を行うためのソフトウェアが開発され，そして公開されている。JASP（JASP Team, 2018；Marsman & Wagenmakers, 2017；Wagenmakers et al., 2018）は，グラフィカル・ユーザーインターフェースを備え，SPSS に類似した操作性を持つフリーソフトウェアである。JASP は，2 節で紹介したような典型的な応用場面について，理解しやすい図や各種情報とともに，規定事前分布を用いたベイズファクターを算出する。また，論文に掲載しやすいようアメリカ心理学会のマニュアルに沿った図表や，R のソースコードを出力することもできる。現在も開発途上のプロジェクトではあるが，JASP は利用の敷居が低く，かつ本格的な利用に耐える特徴を備えている。本論文の図 1，図 2 も JASP の出力を利用したものである。

R のパッケージでは，同様に典型的な場面における既定事前分布を用いたベイズファクターを算出するために，BayesFactor パッケージ（Morey, Rouder, & Jamil, 2015）が利用できる。また，より汎用的に，任意の統計モデルにおける MCMC 推定から得られた乱数列を入力として，4.3 節で紹介したブリッジサンプリングを用いたモデル比較を行うためには bridgesampling パッケージが利用できる。

Savage-Dickey 法を用いたベイズファクターの算出には，MCMC サンプリングを行うソフトウェアである Stan（Carpenter et al., 2017；Stan Development Team, 2017）や JAGS（Plummer, 2003）を R などの統合環境から呼び，その結果得られる乱数列に対してカーネル密度推定を行うことで確率密度値を推定することができる。これらのソフトウェアは，ネストしていないモデル間の比較のために，bridgesampling パッケージと組み合わせて数値的にベイズファクターを算出するためにも利用が可能である。

このように，ベイズファクターを実践的に利用するための機運は高まっており，またそのための道具立ても整いつつある。ベイズファクターは，理論や仮説を表現する統計モデルについて，ほかのモデルと比較しながら，データがどれだけの支持を与えるかを明らかにする。もちろん，統計学的なモデル選択・評価が決して容易な問題でない

ことは，我々は常に意識しておかなければならない。しかし，データの持つ証拠の大きさを定量化し，仮説やモデルを評価し，そして改善につなげていくことは，これまでもこれからも心理学研究の重要な要素であると考えられる。R. A. Fisher と同時代の研究者である H. Jeffreys の提唱によるベイズファクターは，80 年を超える時を経た現代において新たな注目を集めており，今後の心理学研究でさらなる活用が期待される。

引用文献

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis*, *7*, 503–532.

Bayarri, M. J., Berger, J. O., Forte, A., & Garcia-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, *40*, 1550–1577.

Bayarri, M. J., & Garcia-Donato, G. (2014). Extending conventional priors for testing general hypotheses in linear models, *Biometrika*, *94*, 135–152.

Berger, J. (2006a). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 385–402.

Berger, J. O. (2006b). Bayes factors. In S. Kotz, C. B. Read, N. Balakrishnan, & B. Vidakovic (Eds.), *Encyclopedia of Statistical Sciences*. New York, NY: Wiley.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*. doi:10.18637/jss.v076.i01

Chow, S.-M., & Hoijtink, H. (2017). Bayesian estimation and modeling: Editorial to the second special issue on Bayesian data analysis. *Psychological Methods*, *22*, 609–615.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, *1*, 204–223.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, shar hypotheses about chances, the order of a Markov chain. *Annals of Mathematical Statistics*, *41*, 214–226.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York, NY: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Dienes, Z., Coulton, S., & Heather, N. (2018). Using Bayes factors to evaluate evidence for no effect: Examples from the SIPS project. *Addiction*, *113*, 240–246.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34.

Fruhwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal*, *7*, 143–167.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approach to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, *13*, 163–185.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–760.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. DeGroot, D. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics, vol. 2* (pp. 249–270). Amsterdam, the Netherlands: North Holland.

Good, I. J. (1995). When batterer turns murderer. *Nature*, *375*, 541.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). bridgesampling: An R package for estimating normalizing constants. *arXiv Preprint*, arxiv:1710.08162

南風原朝和（2011）量的研究法　東京大学出版会.

南風原朝和（2014）続・心理統計学の基礎　有斐閣.

Hoijtink, H., & Chow, S.-M. (2017). Bayesian hypothesis testing: Editorial to the special issue on Bayesian data analysis. *Psychological Methods*, *22*, 211–216.

伊庭幸人（編）（2018）ベイズモデリングの世界　岩波書店.

JASP Team (2018). JASP (version 0.8.5). [Computer Software]

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*, 203–222.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jeffreys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Jenkinson, D. (2005). The elicitation of probabilities: A review of the statistical literature. *BEEP working paper*, University of Sheffield, UK.

Jones, G., & Johnson, W. O. (2014). Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *American Statistician*, *68*, 42–51.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (2nd ed.). London, UK: Academic Press.（前田和寛・小杉考司（監訳）（2017）ベイズ統計モデリング：R, JAGS, Stanによるチュートリアル　共立出版）

国里愛彦（2018）臨床心理学と認知モデリング　心理学評論，*61*, 55–66.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.（井関龍太（訳）（2017）ベイズ統計で実践モデリング：認知モデルのトレーニング　北大路書房）

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *American Statistician*, *60*, 213–223.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*, 545–555.

松浦健太郎（2016）StanとRでベイズ統計モデリング　共立出版.

Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics*, *22*, 1142–1160.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

三浦麻子（2017）なるほど！心理学研究法　北大路書房.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). R Packages BayesFactor. [Computer Software]

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments." *Journal of Mathematical Psychology*, *72*, 1–5.

村井潤一郎・橋本貴充（2017）心理学のためのサンプルサイズ設計入門　講談社.

村井潤一郎・橋本貴充（2018）統計的仮説検定を用いる心理学研究におけるサンプルイズ設計　心理学評論，*61*, 116–136.

中村國則（2018）高次認知研究におけるベイズ的アプローチ　心理学評論，*61*, 67–85.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*, 125–139.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood nootstrap. *Journal of the Royal Statistical Society, Series B*, *56*, 3–48.

大久保街亜（2016）帰無仮説検定と再現可能性　心理学評論，*59*, 57–67.

大久保街亜・岡田謙介（2012）伝えるための心理統計：効果量・信頼区間・検定力　勁草書房.

Okada, K., Vandekerckhove, J., & Lee, M. D. (2018). Modeling when people quit: Bayesian censored geometric models with hierarchical and latent-mixture extensions. *Behavior Research Methods*, *50*, 406–415.

奥村晴彦・牧山幸史・瓜生真也（2018）Rで楽しむベイズ統計入門　技術評論社.

Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics and Data Analysis*, *54*, 3269–3288.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In D. K. Dey & C. R. Rao (Eds.), *Handbook of statistics 25: Bayesian thinking, modeling and computation* (pp. 115–150). Amsterdam, the Netherlands: Elsevier.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 20–22.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339.

繁桝算男（1995）意思決定の認知統計学　朝倉書店.

清水裕士（2018）心理学におけるベイズ統計モデリン

グ　心理学評論，*61*, 22–41.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Smith, A. F. M., & Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, *42*, 213–220.

Stan Development Team. (2017). *Stan Modeling Language User's Guide and Reference Manual*. [Computer Software]

竹林由武（2018）しなやかな臨床試験デザイン：適応型デザインによる効率化　心理学評論，*61*, 86–100.

竹澤正哲（2018）心理学におけるモデリングの必要性　心理学評論，*61*, 42–54.

豊田秀樹（2015）基礎からのベイズ統計学：ハミルトニアンモンテカルロ法による実践的入門　朝倉書店.

豊田秀樹（2017）実践ベイズモデリング：解析技法と認知モデル　朝倉書店.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology*. Oxford, UK: Oxford University Press.

van de Schoot, R., Schalken, N., & Olff, M. (2017). Systematic search of Bayesian statistics in the field of psychotraumatology. *European Journal of Psychotraumatology*, *8*, 1314782.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of teh Savage-Dickey density ratio. *Journal of the American Statistical Association*, *90*, 614–618.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ..., Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, *25*, 58–76.

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*, 413–426.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, *16*, 117–186.

Wang, L., & Meng, X.-L. (2016). Warp bridge sampling: The next generation. *arXiv Preprint*, arXiv:1609.07690

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*, 752–760.

Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604.

Zellner, A., & Siow, A. (1980). Posterior odds ratio for selected regression hypothesis. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics*, (pp. 585–603) Valencia, Spain: Valencia University Press.

― 2018. 2. 4　受稿，2018. 3. 7　受理 ―