Commentary

ベイズ統計を用いた有意性検定からの脱却

小杉考司
専修大学人間科学部 教授

本稿ではまず,頻度主義的発想な有意性検定の問題点として,p値の分かりにくさ,脆さに加え,理論的仮定が実践上は満たされていないことなどの問題点があることを再確認する。その上で解決策の一つとして,ベイズ統計によるアプローチをあげ,その利点について論じる。ベイズ統計学ではp値のような仮想空間上の値ではなく,実際のデータに基づいた確率分布が得られる。また,この事後分布の導出には補正などが必要なく,データとモデルに基づいて定まるため,検定の多重生の問題などを回避することができる。特に閾値によるカテゴリカルな判断を行わないことが研究実践の観点から重要な利点であることを主張する。最後にこれらの分析方法を実践するための分析環境についても言及する。

1 はじめに

“Nature”のコメンタリー論文で,「統計的有意はやめにしよう」という呼びかけがなされ,これに800人以上の賛同の署名が集まったという(Amrhein, Greenland, and McShane, 2019)。帰無仮説有意性検定(Null Hypothesis Significance Testing, 以下NHST),特にp値の使用についてはアメリカ統計協会も2016年に声明を出しており(Wasserstein and Lazar, 2016),そこではp値にのみ基づいた判断がいかに問題であるかが論じられている。にもかかわらず,いまだに数多くの研究がNHSTの世界から脱却できないでいる。

帰無仮説検定のロジックは,正しく使えば統計の非専門家であっても一定の価値判断ができるという意味で,とても優れた技術である。だからこそ多くの研究領域で用いられ,統計ソフトのパッケージに組み込まれて広まったのだろう。ここで見過ごされがちであるのは,この前提である「正しく使う」ことが,実はことさら難しい技術だったということである。

本稿では,NHSTやp値の何が問題であったのかを改めてまとめ直した上で,これらの問題を克服するためのベイズ統計学の利用について解説を行う。

2 NHSTは何が悪いか

2.1 p値はわかりにくい

NHSTの手続きについて改めて確認しておこう。研究者は帰無仮説と対立仮説を用意する。帰無仮説は研究者の主張とは逆で,差がないとか相関がないといった仮説である。このもっとも保守的な仮説のもと,実際のデータから計算される統計量が算出される確率(p値)が,事前に定めた有意水準よりも低いようであれば,「差がないという仮定が間違っている」と判断する,というものである。

分散分析やt検定など,初めてNHSTを学ぶ多くの人は,この手続きの複雑さを受け入れるのに悩まされる。差があることを証明するために,逆の立場に立ってその仮定を反駁する,いわゆる背理法を用いるからである。さらに,ここから得られる結論は,「差がないと

は言えない」であり，しかもその判断にも二種類の間違い方が生じる可能性がある（Type I/II Error）。また，結論の導出には，「本当は差がないのに，差がないとは言えないと判断してしまう可能性（p値）は5%なので十分に低い」，だから「統計的には意味がある」とする。このロジックは少し飛躍があり，統計学に対する異なる思想が混在している現状（Sober, 2008）に気づかぬままのユーザーも少なくない。

これらの判断の目安とされるp値についても，多くの誤解がある。最もよくあるのが，「p値が小さければ小さいほど仮説が強く支持された」と判断してしまうことである。帰無仮説や対立仮説の正しさは，p値で表されるものではない。p値は帰無仮説が正しいと想定された世界において算出される値，すなわち仮想世界の値であって，現実的データにおける仮説の正しさを反映するものではない。

そしてこの仮想世界は，研究者がどのような世界を前提としているかによって変わり得るものである。前提が変われば同じデータであっても，有意になったりならなかったりする。Kruschke（2015）はこの点を強く批判し，「研究者の意図によって結果が変わるような検証手法は適切ではない」と論じている。ここで言う「意図」をより正確に言うならば，統計モデル，すなわち尤度と事前分布のことである。NHSTのこれまでのやり方では，これらは暗示されているだけで，不十分な言語的記述から読み取るしかなかったのである[1)]。

また，p値は仮説の正しさを表す数字ではなく，ただの目安にすぎない。人文社会科学系では，判断基準である有意水準を5%にすることが一般的であるが，これはあくまでも慣例であり，p=0.06なので「有意な傾向がある」という表現は無意味である。4月2日生まれの人が，「もう少し誕生が早ければ1つ上の学年であったのに」と悔しがっても現実が変わるものではないし，「上級生になる傾向」など存在しないことと同じである。ひいては，「5%水準が緩いから厳しくすれば科学的な精度が上がる」というのも同様に無意味であり，例えば0.01%水準を学会基準としても本質的な問題は解決しない（Trafimow et al., 2018）。

2.2 p値は脆い

アメリカ心理学会のPublication Manualは第6版から，NHSTの結果報告に際して効果量と信頼区間の記載を義務付けた（American Psychological Association, 2010）。これはNHSTの結果の報告が有意であったかなかったか，という1bit判断にしかなっていないことを問題視し，どの程度意味があったのか，という量的な判断についての情報提供を促すものであった。

有意であったかどうかだけが重要であるなら，研究者は有意にするための努力をしてしまいかねない。逆に言えば，少しでも有意でなくなってしまえば，その論文に価値がないと判断してしまいかねない。これは杞憂ではなく，実際にこのような誤った価値観が，心理学においては再現性の問題（友永・三浦・針生,2016; 池田・平石, 2016）を引き起こしたとして批判されている。学会誌に論文が掲載されることだけが目的になり，科学としての知識の積み重ねを放棄するような振る舞いを許してしまう一因として，p値に基づく判断が指摘されている。

検定統計量はサンプルサイズに依存した値であり，p値はサンプルサイズに応じて値が変化する。つまり，本当に差があるかどうかにかかわらず，サンプルサイズが大きければ微小な差異を検出し，あるかないかの判断で言えば，有意差があるという判断をしがちである。「思ったような結果が出なかったので，すこし頑張ってサンプルサイズを増やしたら，有意な差が得られた」というのは美談ではなく，結果の捏造に近いQRPs（Questionable Research Practices, 疑わしい研究実践）である。p値は

こうした脆弱性があるため,サンプルサイズに依存しない効果量（effect size）の記載は今や必須である。

また,p値の判断基準，危険率（α）は，水準を一定に保つために常に補正することを考えておかなければならない。5%の確率で有意というのは5%の間違った判断をする可能性を含んだ評価であるが，この基準での検定をn回繰り返すと,どこかで間違えた判断をする確率は（$1\text{-}0.95^n$）になる。ところが実際の運用では,t検定を行う前に分散の等質性の検定を行ったり,分散分析をした後で下位検定を行ったりすることが少なくない。もちろん分散分析における下位検定では，こうしたαの上昇を抑えるための様々な工夫がなされているが，それが有効なのは1つのデータに対する1回の分散分析の中においてであって,1つの論文の中に複数の分散分析が行われている場合は，どこかで間違った判断を起こしている可能性がある。本来なら,複数の従属変数について考えたい場合は,多変量分散分析によって従属変数を要約し,検定の回数を減らすなどの対策を取るべきだが,この手法は分散分析ほど広まっていない。

このように,p値には目安としての便利さはあるが,実際問題として誤用や誤解を生みがちな問題点を多く含んでいる。これはもちろん，p値やNHSTのロジックが間違っていることを意味するのではなく，問題はp値を基準にして差が「ある」・「ない」といったカテゴリカルな判断をしてしまうユーザーの側にある。ある一定の基準で効果が「ある」・「ない」と離散的に変化するのではなく，検出できるかどうか，効果が大きいかどうかを判断しなければならない。また，NHSTの考え方では，帰無仮説を採択することはできない。差がないことを証明したい場合は，対立仮説に「差がない」をおいたとしても,「差がある」という帰無仮説がいかようにも立てられるので,ロジックが成立しないのである。総じて,「ある」か「ない」かという判断に落とし込もうとすることこそ，考え直す必要がある点である。

ここまで,昨今様々な方面で議論されているNHSTやp値についての問題点を概観した。では,これらの問題を克服する方法がないのかといえば，幸い答えは「ある」であり，その1つがベイズ統計を用いた考え方なのである。

3 ベイズ統計によるブレイクスルー

3.1事後分布はわかりやすい

ベイズ統計学は18世紀ごろ，Thomas Bayes師によって見出された逆確率の法則，つまり現在のデータから過去の状態を推論する方法についての数学的理論に依っている。その後200年の時を待っていま注目を集めている理由は，それまで原理的に可能でも実質的に計算不能だった問題が,計算機科学の発展によって解決されたからである。

ベイズの定理は,モデルのパラメータを$\boldsymbol{\theta}$,データを$\boldsymbol{x}$とし，あるパラメータのもとでデータが得られる条件付き確率を$P(x \mid \theta)$と表すと，

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)}$$

という式で表現される。ここでいうパラメータとは，確率分布関数の形状や位置を作る値のことであるが，より具体的には我々が知りたい未知の値のことだと理解すればよい。例えば回帰分析では切片や傾きの値が未知の値であり，t検定では二群の平均の差がそれである。ベイズ統計学では,これらの値もふくめて,「わからないこと」を確率分布で表現することが特徴的である。この数式の中に含まれている$P(\theta \mid x)$,や$P(\theta)$はいずれも確率分布である。$P(\theta)$は事前確率（事前分布）,データxのもとでのパラメータ$\boldsymbol{\theta}$が得られる条件付き確率$P(\theta \mid x)$は事後確率（事後分布）と呼ばれる。

ベイズの定理の妙は,データを取るまではわ

からなかったパラメータθの値が,データを取ることで事後分布の形に変わることにある。$P(x \mid \theta)$は尤度とよばれるが,調査や実験によってデータは手元にあるので,これはパラメータについての関数である。あるパラメータからこのデータが得られる確率の関数について,事前分布を掛け合わせることで,パラメータがとると思われる値の分布（事後分布）が得られる（事後分布∝尤度×事前分布）。

例えばt検定をベイズ的に行いたい場合は,二群の平均値の差を未知パラメータδのように定める。データはt検定と同じく正規分布からの無作為サンプルを仮定すると,一方の群は$x_A \sim Normal(\mu, \sigma^2)$,他方の群は$x_B \sim Normal(\mu+\delta, \sigma^2)$と表現できる。ここで「~」の記号は,左の変数が右の確率分布に従うことを意味する。μ, δ, σは未知数なので,それぞれに事前分布を考える必要がある。尤度は正規分布関数を適用し,データと組み合わせて考えることで未知数の事後分布を得る。μ, δ, σについての情報がある確率分布として得られるのである。

このように,全てが分布として表現されているため,差の有無のようなカテゴリカルな情報ではなく,差の大きさを最初から捉えられているところが秘めたる利点である。効果を考えたい場合はδの大きさに基づいて考察するのであり,効果の大きさを分布の代表値で点推定してもよいが,確信区間（Credible Intervals）と呼ばれる区間幅で「ここからここまでの範囲の値を取りうる確率がp%」と表現してもよい。信頼区間（Confidential Intervals）と違って確信区間は分布の情報であり,差の大きさを直接確率で表現しているのである[2)]。このように,事後分布の解釈は直感的でわかりやすく,二値判断に陥りにくい性質をそもそも持っている。

3.2事後分布は脆くない

ベイズ統計の方法は,ベイズの基本的なルール,すなわち「事前分布と尤度を考え,事後分布へと情報をアップデートしたものを適用する」というシンプルなものである。対象ごとに統計量や仮定が異なったり,事前の検定,検定,事後の検定を繰り返し適用したりする必要はない。事前分布とデータから事後分布を得る,これだけである。すなわち,検定の多重性の問題に対して頑健である。

事前の検定として,例えばt検定ではLeveneの検定によって分散が均質かどうかを検証する。この帰無仮説は「二群の分散は均質である」であり,危険率によってカテゴリカルな判断を下す。さらにこの場合は,帰無仮説が棄却されたくないことが多く,5%より大きなp値であればよいとされる。もし帰無仮説が棄却されれば,以後のt検定は自由度の補正を行う必要がある。また,例えば反復測定の分散分析では,球面性の検定を行い,分散共分散行列の均質さを検定する必要がある。ここでも均質でないことが示されれば,以後の分析で自由度の補正を行う。このように繰り返される検定と自由度の補正は,NHSTの想定するp値が得られる仮想空間を維持するために必要な操作なのである。

同様のこと,例えばt検定をベイズ的に行う場合,先ほどの記号を使っていえば$x_A \sim Normal(\mu, \sigma_A^2)$,$x_B \sim Normal(\mu+\delta, \sigma_B^2)$とすればよい。すなわち,2つの群で分散が均質でなければ,それぞれ異なる分散σ_A^2, σ_B^2として,それぞれの事後分布を得ればよいのである。球面性の検定の場合も同じで,多次元正規分布の分散共分行列に検定を成立させるための制約を考えるのではなく,分散共分散行列そのものを未知のものとして推定すればよいのである。推定結果に補正はもちろん必要なく,差の大きさについての事後分布をそのまま判断すればよい。

事後の検定についても同様である。分散分析によって差が見られたので,次にどこに差が

あるかを見つけるために,危険率 α を調整しながら主効果, 交互作用を順次検討し, 結果を総合的に解釈するというのは大変煩雑な作業である。ベイズ的な分析の場合, どのような実験デザインであれ,データと尤度, 事前分布が定まれば事後分布が得られる。この事後分布は1つの分布（複数のパラメータによる同時確率分布）であるから, どの側面から切り取って判断しようとも結果は変化しない。危険率の調整や細切れになった結果を総合的に解釈する必要もない。例えば2×3の分散分析で交互作用が見出され, μ_{12}とμ_{13}の間の差が見たいのであれば, $\delta = \mu_{12} - \mu_{13}$を算出してその分布を検討するだけでよいのである。

4 ベイズ分析への第一歩 パッケージの利用

ここまで述べてきたように,ベイズ統計分析を行う利点はとても多い。なかでも検定結果に誤用や誤解が入りにくいため,むしろ初学者はベイズ統計から始めるべきであるというのが筆者の主張である。これまでベイズ統計学が主流にならなかったのは,コンピュータによる計算速度が不足していたからに過ぎないのである。

平均の差の検定は, それが二群であろうと要因計画であろうと,線形モデルとして表現できる（一般線形モデル）。このように形式化されれば,ベイズ的な分析であるかどうかは実は係数（未知数）の推定方法の違いにすぎないとも言える。従来は最尤法による点推定であり, その結果を使ってカテゴリカルな判断を下していたのだが, これをベイズ推定による事後分布に基づく判断に変えるだけで,ベイズ分析の利点を全て享受できる。

モデルが確定しているため,ベイズ統計の中でも分析ツールのパッケージ化は進んでいる。例えば統計環境Rではbrms（Bürkner, 2017 ; Bürkner, 2018）やrstanarm（Goodrich et al. , 2018; S. Brilleman et al.,2018）といった一般化線形モデルのベイズ推定パッケージが用意されている。特にbrmsはStanという確率的プログラミング言語（Carpenter et al.,2017）の開発チームによるパッケージであり, とても使い勝手がよい。説明変数をカテゴリカル変数（factor型）であることを明示して回帰分析を行うと, その結果は各群の平均値の事後分布を確信区間とともに示す形となる。このような形式で結果を得ることは,カテゴリカルな判断に陥らずに豊富な情報を得ることになる。

他にもJASPとよばれる統計解析のフリーソフトウェアがある（JASP Team,2018）。JASPはGUIを使って操作が可能であり, 従来の分析とベイズ的分析がメニューの同じグループに配置されているので, 両者の結果を比較検討することもできる。平均値の差の検定については,$\boldsymbol{p}$値ではなくベイズファクター（BF）と呼ばれる指標とともに表示される。BFは2つの仮説（モデル）について, 一方が他方に比べてどの程度データに支持されているかを示している指標である。BFを使う利点の第一は,$\boldsymbol{p}$値と違ってサンプルサイズを増やす途中で判断が変わることがないことである。第二に,「差がない」というモデル($\mu = 0$)にくらべて「差がある($\mu \neq 0$)」というモデルがどの程度データに支持されているか, といった比較ができるため,「差がない」というモデルを積極的に採用できる点が挙げられる。

これらの新しいツールを使うことで,今後はt検定や分散分析といった手法が一新されることは論をまたないであろう。

5 まとめと展望

本稿では, NHSTの問題点を指摘し, それを克服するためのベイズ統計学の利用について解説を行った。

NHSTの問題として,$\boldsymbol{p}$値の複雑さ, 難解さ,

誤用のしやすさ，そもそもの限界などを指摘したが，分散分析をもはや使うべきではない別の理由も存在する。t検定や分散分析といった一般線形モデルは，パッケージングできる程度に単純な古典的モデルであり，「これまでの調査や実験などの研究はそのモデルに合致するように研究デザインを工夫せざるを得なかった」という限界に覆われているのである。「古典的」と表現したことから明らかなように，現在ではより進んだ線形モデルが存在する。

一般線形モデルは正規分布に従うデータに限定されていたが，実際のデータは必ずしも左右対称の単峰分布とは限らない。例えば世帯収入などの分布は対数正規分布に従うことが知られており，このデータに対して平均値を検定の対象にするのは適切ではない。こうした分布の問題に対応し，様々な分布の平均パラメータを線形モデリングできるようにしたのが一般化線形モデル[3]である。本稿でも紹介したbrmsやrstanarmは一般化線形モデルに対応しており，一般線形モデルをその特殊系（ガウス分布モデル）として含んでいる。

また，調査データが複数の群や性質によって異なっているが，それが混合した形で得られていることもあるだろう。そのような場合は，一般化線形混合モデルで対応しなければならない。さらに反復測定や階層化されたデータに対しては，階層線形モデルなどが考えられ，こうした複雑なモデルになってくると，最尤推定は現実的に不可能で，ベイズ推定を行うしかないのが現状である。とはいえ，これらも統計パッケージで対応できる問題であり，要は柔軟なモデルの適用を考えていくと，推定法としてのベイズが唯一の現実的な回答なのである。

もちろん，これまでどおりの分析モデルで，推定法をベイズに変えるだけでもカテゴリカルな判断から自然と脱却でき，その恩恵は少なくない。p値0.05のような仮想空間の恣意的な目安にとらわれず，実質的な差に目を向けるという，研究者の本質にたちかえればよいのである。

ただし，この推定法はもっと強力で，線形ではないモデリングにおいてもパラメータの推定を可能にする。研究者はより細かなモデリングで現象の記述ができ，より実質的，具体的な問題を検討できるようになるだろう。いずれにせよ，もはや分析モデルの要請によってデータの形を変えたり，研究をデザインする必要はないのである。不自由な世界から脱却し，自由なモデリングの世界へと続く道は，ベイズ統計学で舗装されている。

注

1) Kruschke（2015）のいう研究者の意図と尤度については，小杉（2019）も参照のこと。

2) 確信区間も信頼区間もいずれもCIで表されるのは誤解を招きかねない表現ではあるが，同じ略記であっても意味が違うというベイジアンのユーモアでもあるだろう。

3) もう一つの方法として，平均値を目的とした回帰係数の最適化を行うのではなく，中央値など任意のパーセンタイル点に対するモデルであるQuantaile Regression法というのもある。詳しくは（Hao, Naiman, and Naiman, 2007）などを参照。

文献

American Psychological Association, 2010, *Publication Manual of the American Psychological Association*. 6th ed. Washington, D.C.: American Psychological Association.

Amrhein, Valentin, Sander Greenland, and Blake McShane, 2019, "Scientists Rise up Against Statistical Significance." *Nature,* 567：305-307.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe, 2018, "Joint Longitudinal and Time-to-Event Models via Stan." (https://github.com/stan-dev/stancon_talks/).

Bürkner, Paul-Christian, 2017, "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28.

————, 2018, "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell, 2017, "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76 (1): 1–32.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman, 2018, "Rstanarm: Bayesian Applied Regression Modeling via Stan." http://mc-stan.org/. (http://mc-stan.org/)

Hao, Lingxin, Daniel Q Naiman, and Daniel Q Naiman, 2007, *Quantile Regression*. 149. Sage.

池田功毅・平石界, 2016,「心理学における再現可能性危機──問題の構造と解決策」『心理学評論』59 (1): 3–14.

JASP Team, 2018, "JASP (Version 0.9)[Computer software]." https://jasp-stats.org/.

小杉考司 , 2019,「新しい統計学とのつきあい方」『基礎心理学研究』37 (2): 167–73.

Kruschke, John K, 2015, *Doing Bayesian Data Analysis, a Tutorial with R,JAGS and Stan*. 2nd ed. NewYork: Elsevier Inc. (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)

Sober, Elliot. 2008. *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press. (松王政浩訳 ,2012,『科学と証拠──統計の哲学入門』,名古屋大学出版会)

Trafimow, David, Valentin Amrhein, Corson N. Areshenkoff, Carlos J. Barrera-Causil, Eric J. Beh, Yusuf K. Bilgiç, Roser Bono, et al. 2018. "Manipulating the Alpha Level Cannot Cure Significance Testing." *Frontiers in Psychology* 9: 699.

友永雅己・三浦麻子・針生悦子 , 2016,「心理学の再現可能性──我々はどこから来たのか我々は何者か　我々はどこへ行くのか」『心理学評論』59 (1): 1–2.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The Asa's Statement on P-Values: Context, Process, and Purpose." *The American Statistician* 70 (2). Taylor & Francis: 129–33.