研究ノート

# 教育統計における課題と今後の展開

## ―効果量とベイズ統計とその応用から―

阿　部　敬　信

Isseues and outlook for the future in educational statistics:
About effect size and application of Bayesian statistics

Takanobu ABE

**【要　　旨】**

本稿では，まず教育統計で課題となっている帰無仮説の棄却の問題や近年話題となっている効果量を巡る問題，さらに有意差に係る統計的確率論的な意味について取り上げた後に，従来の Neyman and Pearson が確立した統計学に対して，急速に広まりつつあるベイズ統計の基本公式について概説し，ベイズ統計の教育統計への応用について，筆者の研究からその構想について述べる。ベイズ統計については教育統計に限らず，統計に関連する諸分野で，これからの応用が期待されることを示した。

**【キーワード】**

教育統計　　帰無仮説　　標本サイズ　　効果量　　ベイズ統計

## 1　教育統計における課題

### (1)　帰無仮説の棄却

教育統計においては，分布の正規性を前提にした統計的仮説検定がほぼ絶対的な位置を確立しているといってもよい。実証的研究においては，まず確実に仮説検定結果のいわゆる p 値が記されており，それにより帰無仮説の採択・棄却によって，結論が付けられていることが大部分である[1)]。しかしながら実際には「帰無仮説が正しくなければ，対立仮説は正しい」あるいはその対偶である「対立仮説が正しくなければ帰無仮説が正しい」は実際には成立していない。

まず，p 値は有意水準 $\alpha$ を事前に決定しておいて，次のように判断する。

$p \geqq \alpha$：帰無仮説は棄却できない　　(1)

$p < \alpha$：帰無仮説は棄却できる　　(2)

この有意水準 $\alpha$ は慣例としてし $\alpha = 0.05$若しくは $\alpha = 0.01$として決められている。これは「慣例」でしかない。ある意味，恣意的に決められているともいえる。この意味は100回試行して５回しか生起しないのであれば，ほとんど生起しないと考えられるからという確率論的な確からしさが背景にあるのみである。

次に，(1)の場合について考えてみる。ここではあえて「帰無仮説は棄却できない」と記した。なぜなら(2)の「帰無仮説が棄却できる」に対して採択されるのは実際には対立仮説であるからである。この意味は，つまり「帰無仮説が採択される」ではなく，帰無仮説と対立仮説のどちらとも採択できるとも採択できないともいえない，つまり判断を保留していることに他ならないということである[2)]。

### (2) 標本サイズと効果量

さらに標本サイズの問題がある。例えば次のような例題を考える。

> A市とB市の20歳代の人を無作為に1000人ずつ抽出して、ある知能検査を実施しました。その結果、どちらの分布も正規分布に従っていて、平均と標準偏差が次のようになりました。
>
> A市　平均：110　標準偏差：10
>
> B市　平均：109　標準偏差：10
>
> A市とB市の20歳代のそれぞれ1000人の知能指数の平均に相違があるかを適切な統計的検定で調べてください。

ここで用いる検定は，当然，対応のない２群のt検定である。結果はt（1999）＝2.23　p＜.05となり２群の平均には５％有意水準で有意差が認められる。ここで有意差が認められてしまったのは，標本サイズの大きさが効いてしまっているのは明らかである。２群の平均の差の絶対値に対して標本サイズが大きすぎるから，本来であれば有意差が認められないケースであるにもかかわらず，有意差が出てしまったといえる。逆にいえば，標本サイズを大きくしていけば，p値を限りなく小さくできてしまうということである[3)]。

近年は，このようなケースを検討できるようにするために効果量も積極的に報告することが求められるようになってきている[4)]。効果量はp値や検定統計量とは異なり，帰無仮説が正しくない程度を量的に表現できる。さきほどの例題で見たように一般に検定統計量は標本サイズに影響を受けるが，効果量は標本サイズを大きくすると，推定の精度を上昇させることはあっても効果量そのものへの影響はない。

例えば対応のない２群のt検定における効果量gは次のようになる[5)]。なお，２群のそれぞれの標本平均を$m_1$，$m_2$とし，両群に共通な母標準偏差をsとする。ここで$s_1^2$，$s_2^2$はそれぞれの群の不偏分散である。

$$s=\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}} \tag{3}$$

$$g=\frac{m_1-m_2}{s} \tag{4}$$

なお，対応のない２群のt検定における検定統計量tと効果量gの間には，次のような関係がある。

$$t=g\sqrt{\frac{n_1n_2}{n_1+n_2}} \tag{5}$$

先の例題の効果量gを(5)の式によって求めると，g＝0.10となる。Cohenの基準[6)]によればg＝0.2は「小さな効果量」であり，例題の結果はさらに効果量が小さくなっている。効果量が小さいにもかかわらず有意差が得られている例題はやはり標本サイズが大きいために本来であれば有意でない場合が有意となっている第１種の過誤を犯している可能性が高くなる。

これを防ぐためには，検定力分析を行い，十分な大きさの効果量を得ることができるような標本サイズをデータを得る前に求めておいてから，研究デザインと標本の選定を行う必要がある。しかしながら，近年の教育現場においては実際に研究を行うフィールドを確定することすら困難な状況があり，研究期間も限られる中で，研究デザインの段階で十分に標本サイズを考慮することは難しいというのも現実である。

### (3) 統計的検定の確率論的意味

標本の選定では無作為抽出が大原則であり，この前提がないとそもそも正規性を前提にした区間推定などできない。しかしながら，実際問

題は，先にも述べたように，標本の選定ですらさまざま制約があり困難な状況があり，ここに研究者の恣意性が入る余地がある。

つまり，統計的検定は一見定められた条件と手続きによって客観的に検定統計量を算出しているように見えて，その実は研究者の恣意性が入る余地が多いのである。しかも，客観的と思える検定統計量ですら，推定値であり，確率的現象を表しているものにすぎないのである。例えば慣例として用いられており，何ら根拠のない５％有意水準であっても，100回の試行の中で５回しか生起しないということは滅多に生じないことだから，棄却してもいいのではということでしかない。逆に言えば，５回は生起するということなのである。

何度も強調しているが，統計的検定は、１回の実験や調査において、さまざまな要因に関して限られた条件下でしか検討を行うことができない。したがって、１回の実験や調査における仮説の支持は、あくまでも仮説を支持する一つの例を示したにすぎない。その限界をよく考えておかねばならないのである[7]。

### ⑷ 今後の教育評価

そのために，教育統計では多面的な評価ということがよくいわれる。児童生徒を偏差値で輪切りにしてレッテルを貼るのではなく，それはあくまでも，児童生徒がもっている個々の特性の一側面にすぎないのだから，さまざまな角度から評価することが大切だといわれている。ポートフォリオ評価や授業中の態度，ノートの記述など，さまざまな情報を収集して多面的な評価が望まれる。その一つが教育統計による数値評価なのである。

## ２ ベイズ統計の教育統計への応用

### ⑴ 頻度論的確率と主観確率

これまで見てきた教育統計の考え方は，従来からの統計学で確立されてきたNeyman and Pearsonの考え方に依拠している。確率変数であるデータの母集団が従う分布をあらかじめ規定して，パラメータとしてその母集団に唯一の固有な値が存在すると仮定して，一定の数学的手続きによって推定していく。その多くは，正規分布を仮定しており，平均値と分散がその母集団のパラメータをしている。これは母集団からのランダムサンプリングを前提として，反復可能な試行において生起する事象の相対頻度を基本としたものであるから，頻度論的確率(frequentist probability）とよばれている。しかしながら，現実の生活においては，人間は反復試行して相対頻度を求めてから確率で判断するのではなく，それまでの経験や思い込みといった主観的な見方によって，おおよそこれぐらいという確率といえば確率といえるような経験値によって判断していることが多い。このような確率的な判断は主観確率（subjective probability）とよばれている。従来の頻度論的確率に基づいたNeyman and Pearsonの統計学に対して，このような主観確率を，仮説の真偽や母数の値に積極的に認めて「データに基づいて仮説を評価する母集団について推論する」統計学のことをベイズ統計学(Bayesian statistics)という[8]。ベイズ統計学はベイズの定理とその定理に基づく統計的推定であるベイズ推定（Beyesian inference）からなっている。

### ⑵ ベイズ統計の基本公式

一般にある事象Aが起こったという条件のもとで，事象Bが起きる確率を，事象Aのもとで事象Bが起きる条件付確率といい，次のように表すことができる。

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{6}$$

ここで，$P(A \cap B)$は事象Aと事象Bが同時に起きる確率である。これを$P(A \cap B)$について解くと，次のようになる。

$$P(A \cap B) = P(B \mid A)P(A) \tag{7}$$

ここで，AとBは交換可能である。

$$P(A \cap B) = P(A \mid B)P(B) \tag{8}$$

これを式(6)に代入すると次のようになる。

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{9}$$

これがベイズの定理である。

ここで，Aを仮説（原因），Bを得られたデータであると考えると，P(A) は原因の確率であり，データBを得る前の確率と捉えることができることから事前確率と考えることができる。

次にP(A | B) はデータBを得ることができる原因がAである確率と考え，事後確率と考えることができる。またP(B | A) は原因Aによりデータが得られる確率であり尤度とよばれる。このように事前確率P(A) が原因Aのときにデータを得る事後確率P(A | B) を求める際に取り込んでいることが特徴であり，主観確率とよばれる所以である。

さらに一般化を図るために，データBを得ることができるいくつかの原因$A_1$，$A_2$，…，$A_n$があるとする。このとき，データBが得られたとすると，その原因が$A_i$である確率は次のようになる。

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots P(B \mid A_n)P(A_n)} \tag{10}$$

式(10)をベイズの展開公式とよぶ。

ここで，ベイズの展開公式をベイズ統計で扱えるようにする。従来の頻度確率論では，固定されていた母数を，ベイズの展開公式における原因Aと考える。母数は連続量であり何らかの分布をしていると考えるのである。母数を$\theta$とすると，事前確率P(A) は確率密度関数$\pi(\theta)$ととらえることができるようになる。(10)の分母は，データを得た段階では，データを得る確率P(B) となるので，定数を見なすことができ，次のようになる。

$$\pi(\theta \mid B) \propto F(B \mid \theta)\pi(\theta) \tag{11}$$

(11)はベイズ統計の基本公式である。つまり，データBが得られたとき，母数$\theta$の事後分布は，母数$\theta$の確率密度関数のもとで，データBが得られる確率（尤度）とデータBを得る前の母数$\theta$の事前分布の積に比例するということである。

例えば，ベイズ統計において事前分布及び事後分布として用いられることが多い分布関数であるベータ分布がある。ベータ分布の確立密度関数は次のように与えられる。

$$F(B \mid \theta) = k\theta^{p-1}(1-x)^{q-1} \tag{12}$$
（kは定数，$0<p<1, 0<q<1$）

なお，ベータ分布においては，平均値$\mu$と分散$\sigma^2$は次のようになる。

$$\mu = \frac{p}{p+q} \tag{13}$$

$$\sigma^2 = \frac{pq}{(p+q)^2(p+q+1)} \tag{14}$$

ここで，次のように尤度$\pi(\theta)$を二項分布とする。

$$\pi(\theta) = {}_nC_r\,\theta^r(1-\theta)^{n-r} \tag{15}$$

尤度$\pi(\theta)$を二項分布とすると，自然な共役事前分布（natural conjugate prior distribution）としてベータ分布を用いることができる。

よって，事後分布は，(11)に代入して，次のように求めることができる。

$$\begin{aligned}\pi(\theta \mid B) &\propto F(B \mid \theta)\pi(\theta) \\ &= {}_nC_r\,\theta^r(1-\theta)^{n-r} \times k\theta^{p-1}(1-\theta)^{q-1} \\ &\propto \theta^{r+p-1}(1-\theta)^{q+n-r-1}\end{aligned} \tag{16}$$

つまり，事後分布もベータ分布となる。ただし，ベータ分布のパラメータが事前分布ではB（p,q）から事後分布ではB（p+r, q+n−r）となっていることに留意する必要がある。

### ⑶ ベイズ統計の考え方と教育統計への応用

このようにベイズ統計は，得られたデータから尤度を算出し，事前分布を仮定して，ベイズ統計の基本公式(11)から事後分布を推定していき，さらにこの事後分布が次のデータの事前分布となってデータの分布に適合させていくことができる。これをベイズ更新（Bayesian update）という。導入の事前分布は確からしいという主観に基づいて一様な分布を用いてもよい。このことを理由不十分の原則（principal of insufficient reason）とよんでいる。しかし，ベイズ更新によって，逐次事前分布は更新されていく。

このようにベイズ統計は，人間の経験や思い込みを統計の処理過程に導入しているところが，これまでの頻度論確率とは決定的に異なっている点である。よって，教育統計への応用は十分に考えることができる。高[9]では，大学生のGPAをデータとして，対象とする学生の母集団の学力分布の分析を行い，一定の合格率を設定した場合の各学生の合格可能性の推定を，ベイズ統計における階層ベイズ法を用いて行っている。共通能力$\beta$と個々の学生の能力$\gamma$をロジットモデルという統計モデルを導入して各学生の成績を算出し，その成績から能力$\gamma$を有する学生人数の確率分布を導出している。これにより一定の合格率による学生の人数を導き出している。一定の合格率を就職試験の合格率とすれば，就職の可能性を確率として示すことができる。

### ⑷ 今後の応用へ向けて

筆者の研究分野で構想しているベイズ統計の応用について述べる。

筆者は現在，日本手話・日本語バイリンガル児童の第二言語としての日本語の読解力評価の研究を行っている。第一言語とは異なり，第二言語となると個々の習得状況の差がかなり見られる。それを一つの手続きによって統一的に評価を行うことはかなり困難な作業である。そこで，最初にいくつかの検査素材となる本を準備しておき，その中から検査水準に乗ることができる本を被検査者である児童が選定し，その理解の実態を何らかの事後テストを第一言語である日本手話による検査教示によって行うという構想で進めている。この被検査者である児童一人一人によって検査の材料が異なることから，統一的な解釈と評定について何らかの補正が必要と考えていたが，それでは膨大な被検査者が必要となる。ここで，母数が固有の値ではなく事前分布として個々の被検査者ごとに与えることできるベイズ統計によって事後分布への統計モデルの適用ができるのではないかと構想している。

次に，筆者がベイズ統計の適用を構想している研究は，幼稚園における幼児の遊びの質を評価していく研究である。「初等中等教育における教育課程の基準等の在り方について（諮問）」[10]によれば，幼児教育の分野においても接続の課題とともに何らかの評価を考えていかねばならない状況にあるようである。実際に第三者評価の記述の充実など幼稚園の特性に応じた学校評価を推進するための「幼稚園における学校評価ガイドライン（平成23年改訂）」[11]においては，「評価項目・指標等を検討する際の視点となる例」として、便宜的に分類した学校運営における12分野ごとに例示がされており，例えば「幼稚園教育要領の内容に沿った幼児の発達に即した指導の状況」を評価することが求められている。これに対して，河邊[12]は「「遊びの質」は絶対的尺度をもって測定可能なものではない」として，「自己充実が深いときに学びが大きいとすれば、遊びの質を問う観点として、すなわち、遊びにおける自己課題（遊び課題）が、人やモノとのかかわりの中で、どのような関係をもって生み出され、遂行されるのかを視野に入れることが重要と言える」として，「モノ・コトとのかかわり」と「他者とのかかわり」という視点からの遊びのプロセスこそが遊びの質と捉えると述べている。時間軸に沿った流れで「遊びの質」といういくつかの母数が連続的に分布しているととらえると，この観測値の分布と事前分布から，事後分布としての遊びの質が数学的に導出できないかと考える。本稿では

触れなかったが，ベイズ統計では階層ベイズモデルにマルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo procedure; MCMC）を用いることで，より複雑な数理モデルを導出できることから，全く無縁と思われた遊びの質の検討，つまり「幼稚園教育要領の内容に沿った幼児の発達に即した指導の状況」の評価に何らかの寄与ができるのではないかと構想している。

## 【引用文献】


1）波田野結花・吉田弘道・岡田謙介，『教育心理学研究』におけるｐ値と効果量による解釈の違い，教育心理学研究，63（２），151－161, 2015.
2）久保拓弥，データ解析のための統計モデリング入門－一般化線形モデル・階層ベイズモデル・MCMC，2012, 岩波書店.
3）前掲書１）
4）石井秀宗・吉田寿夫・岡田謙介・南風原朝和，心理学研究における効果量の活用と報告－APA の指針をふまえて, 教育心理学年報, 52, 234－237, 2013.
5）大久保街亜・岡田謙介，伝えるための心理統計―効果量・信頼区間・検定力，2012, 頸草書房.
6）Cohen,J.,The t-test for means, Statistical power analysis for the behavioral science ２nd edition., 19－74, 1988, L. Erlbaum Associates.
7）吉田寿夫，本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本，1998, 北大路書房.
8）南風原朝和，続・心理統計学の基礎－統合的理解を広げ深める，2014, 有斐閣.
9）高正博，ベイズ統計の応用に関する検討，東海大学開発工学部紀要，20, 147－152, 2010.
10）文部科学省，初等中等教育における教育課程の基準等の在り方（諮問），2014.
11）文部科学省，幼稚園における学校評価ガイドライン（平成23年改訂），2011.
12）河邊貴子，遊びの質を読み取る視点－モノとのかかわり・人とのかかわり，日本保育学会第68回大会要旨集，234, 2015.

※　本稿「２　ベイズ統計の教育統計への応用」の「(2) ベイズ統計の基本公式」は，次の文献を参考にした。
涌井良幸，道具としてのベイズ統計，2009, 日本実業出版社.
松原望，入門ベイズ統計－意思決定の理論と発展，2008, 東京図書.
Gelman, A. et.al., Bayesian Data Analysis ３rd edition, 2013, Chapman and Hall/CRC.


※本研究は，日本学術振興会学術研究助成基金助成金基盤研究Ｃ（日本手話・日本語バイリンガル児童生徒の読解力の向上と評価に関する研究；研究課題番号：15K04582）による研究成果の一部である。