解説

「統計的に有意」で満足していませんか？ ― 統計的帰無仮説検定の問題と対応 ―†

島田　英昭*1・井関　龍太*2

1. はじめに

学生「先生，検定したら，有意になりました！」

教員「おお，やったな！これで論文投稿できるな」

このような会話が，日々全国の研究室で行われているのではないだろうか．論文投稿は研究者や研究室にとって重要な営みである．その前提には「成果」がある．データを取って統計的に検証することが求められる研究領域では，成果を示す一つの道具として「統計的帰無仮説検定」が頻繁に使われる．省略して「統計的仮説検定」「仮説検定」「検定」等の呼び方も使われるが，本稿は以下「検定」と呼ぶことにする．

学生「先生，今回の実験，残念ながら有意になりませんでした」

教員「残念．それだと査読に通らないよなぁ．やり直すか」

このような会話もよくあるのではないだろうか．二つの会話が意味しているのは，検定が論文投稿にとっての「関門」になっていることである．有意であれば喜び，有意でなければ悲しむ．論文掲載が仕事の成果であり，研究者としての評価や若手研究者の就職に直結し，その前提に統計的有意性がある．有意かどうかで感情的になることは，自然なことだろう．

しかし，検定については古くから多くの問題が指摘されている．筆者らの研究領域である心理学では，科学研究にはたいへん重要である再現性の問題が大きく取り上げられて [1]，その原因の一つでもある検定を見直そうという機運が高まり，「統計革命」と言われることもある [2].

筆者らは，心理学の領域で研究を行っている研究者である．統計学の専門家ではなく，統計のユーザーにすぎないが，近年の統計環境の変化に対して，日々の研究で対応を求められている．本稿では，統計ユーザーの立場から，検定の問題と対応について，心理学領域で起こっている統計環境の変化を紹介する．

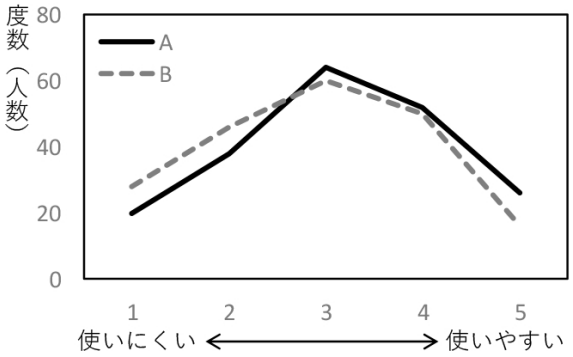


図 1　本稿例の度数折れ線グラフ

2. そもそも統計的帰無仮説検定とは

検定の問題を指摘する前に，検定について簡単に振り返る．我々は統計ユーザーであり，必ずしも数学的な理解を厳密に求められているわけではない．しかし，全くの理解なしに肯定あるいは否定するわけにはいかない．そこで，最低限の理解を共有するために，以下の例（以下，「本稿例」と呼ぶ）を用いて，検定が何をしているのか，復習する．

2.1　本稿例

あるインターネットサービス A を開発し，従来サービスである B と比較し，A の有効性を示す実験をインターネット上で行った．合計 400 名の実験参加者を 200 名ずつランダムに A と B の 2 条件に割り振り，それぞれのサービスを使用してもらった後，サービスの「使いやすさ」について「1. とても使いにくい」から「5. とても使いやすい」の 5 段階評価を求めた．結果を度数折れ線グラフにして，図 1 に示す．A，B の平均値はそれぞれ M=3.13（SD=1.16）, M=2.90（SD=1.16）であった．t 検定により平均値を比較した結果，有意な差がみられた（t(398)=2.25, p<.05，両側）．したがって，サービス A は B に比べ，利用者が感じる使いやすさの点で優れると考えられる．

細かく指摘すれば，本来はこのような段階評価は連続量ではないので平均値の算出に適さない（いわゆる「間隔尺度」ではなく「順序尺度」である）．また，近年では，このタイプの，いわゆる「対応のない t 検定」では，はじめから自由度を調整するウェルチ法を使うことが勧められることが多い [3]．さらに，使いやすさ以外の指標も使って総合的に判断すべきといった議論はあるだろう．しかしここでは，これらの議論は置いて，「2

---

† Are You Satisfied with Statistical Significance? Issues in and Resolutions for Null Hypothesis Significance Testing
Hideaki SHIMADA and Ryuta ISEKI

*1 信州大学学術研究院教育学系
Institute of Education, Shinshu University
*2 大正大学心理社会学部
Department of Human Sciences, Taisho University

群の平均値の比較」を行う場面において平均値の差が有意であったという事実について考えよう．

2.2　統計的帰無仮説検定の意味

統計ユーザーである立場から考えると，この結果から「苦労して作った A の効果が示された！」と考え，学会発表と論文投稿の段階に移るのかもしれない．その前に，一歩立ち止まって，そもそも検定とは何なのか，考えよう．

検定は，数学における背理法に似たものとして理解することができる．背理法が使われる典型例として，命題「$\sqrt{2}$ は無理数である」の証明がある．背理法は，その否定「$\sqrt{2}$ は有理数である」，すなわち「$\sqrt{2}=p/q$（p, q は自然数）と表現できる」を仮定して，その世界の中で矛盾を導き，結果としてはじめの命題が正しいことを示す方法である（詳しくはウェブを検索してほしい）．

本稿例の命題は「2 群の母平均が異なる」である．母平均とは，母集団の平均値である．得られたデータはこの母集団から得られたサンプル（標本）であると考える．そして，この命題の否定が「2 群の母平均が等しい」である．この「2 群の母平均が等しい」という前提のもとで，そのようなことがない，正しくは「ほとんどあり得ない」ということを確率的に示すことで，間接的に「2 群の母平均が異なる」ことを示す．これが検定の論理である．「ほとんどあり得ない」を示す確率としては，一般的に 5% が用いられる．確率が入り込むことが異なるが，背理法の論理を参考にすると理解しやすい．

検定では，もともと示したい命題（差があるという主張）を対立仮説，その命題の否定（差がないという主張）を帰無仮説と呼ぶ．帰無仮説という名前からわかる通り，帰無仮説は結果的に否定されることが期待されている．

ここでは 2 群の平均値の差について議論したが，対立仮説と帰無仮説というレベルで一般化すれば，他のさまざまなタイプの検定，たとえば分散分析，無相関検定，重回帰分析の偏回帰係数や決定係数の有意性検定，クロス集計表の直接確率計算やカイ二乗検定など，論理的にはすべて同じである．以下，さまざまなタイプの検定に一般化できることを前提に，2 群の平均値の比較を例に議論していく．

2.3　サンプルから母集団の性質を推測する論理

検定では，得られたデータはある母集団から抽出されたサンプルであると考える．研究者が知りたいことは母集団の性質であり，2 群の母集団の平均値の差である．これを図 2 に示す．「A の利用者母集団」と言うと，「大学生」「日本人」といった具体的な集団ではないため，理解しにくいかもしれない．ここでは『A を利用した人の「使いやすさ」に対する評定値の集まり』と理解するとよいだろう．実際，ここには{1,2,4,5,3,2,3,3,. . . }といった評定値が格納されている．検定の前提として，得られたサンプルから母集団の性質を推測し，一定の結論を下そうとしている構図を考えることが重要である．

ここで，帰無仮説のもとで，すなわち 2 群の母平均が等しいという前提のもとで，同じサンプリングをくり返すことを考えてみよう（図 3）．実際の研究は 1 度しかデータを取ることができないことが一般的なので，ここでの議論は仮想的なものである．二つの母集団から 200 ずつサンプルを取り出し，それぞれの平均値を計算し，平均値の差を算出することをくり返す．2 群の母平均は等しいはずなので，その平均値の差は 0 になるだろう．しかし，サイコロの目が 3 回連続で同じ目が出ることがあるように，取り出すサンプルには偏りが含まれることがある．そのため，2 群の平均値の差にもズレが生じる．通常，ぴったり 0 になることはなく，時にはプラス，時にはマイナスの数値をとる．通常このズレは小さいが，時には大きく偏る．この偏りの程度を図示したのが，図 3 右側の分布である．帰無仮説のもとでは 2 群の平均値の差はないので，サンプルの多くは中心の 0 に集まるが，まれにズレが大きくなることがある様子をあらわしている．

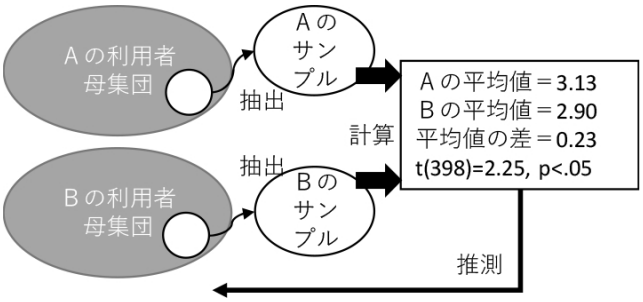


図 2　本稿例の母集団とサンプル

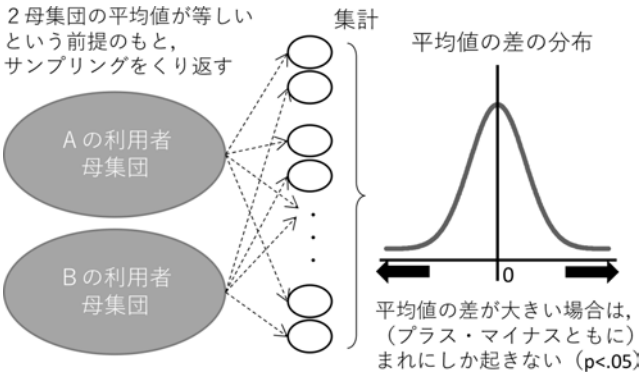


図 3　サンプリングの繰り返しと平均値の差の分布

さて，我々が調査を行った場合には，通常サンプルは 1 回分しか得られていない．もし 2 群の母平均が等しければ，平均値の差は図 3 の分布の 0 に近い位置にあることが期待される．しかし，今，手元のデータをみたところ，このズレが大きかったとしよう．2 群の母平均が等しいとしても，偶然大きなズレが生じることもある．しかし，可能性はもう一つある．それは，2 群の母平均が等しいという，そもそもの前提が誤りである可能性である．2 群の母平均の差がもともと 0 でないのなら，手元のサンプルの平均値の差が 0 からズレていることは大いにあり得るし，むしろ自然なことである．検定では，このズレの大きさについて，「もし帰無仮説が正しいとしたらそのような大きさのズレが生じることがどれくらいありそうにないのか」を確率で表す．この確率を「p 値」と呼ぶ．ズレが大きいほど，言い換えると図 3 右側の分布で平均値の差が中央から離れるほど，p 値は小さくなる．平均値の差が分布の左側，あるいは右

側に大きく偏る時，p 値は 0 に近づく．したがって，p 値が小さいほど，平均差の 0 からのズレが大きいことを意味する．そこで，この p 値が一定の基準値よりも小さい時，「ズレが大きい」と評価すればよいだろう．この「基準値」としては，通常は 5% が用いられる．これは有意水準と呼ばれる．もし，ズレが 5% 未満の確率でしか起きないほど大きいものであったら，2 群の母平均が等しいという前提を否定する．すなわち 2 群の母平均が異なると推測する．これが検定の論理である．

問題は，このズレの分布を定量的に把握できるのかという部分である．本稿例では，t 検定の背景理論がそれを可能にしている．数学的な詳細は省くが，2 群の平均値と標準偏差をそれぞれ算出し，そこからある計算式に入れると t 値と呼ばれる値が得られ，この t 値は t 分布と呼ばれる分布に従うことがわかっている．t 分布は正規分布と同じく数学的に定義される．したがって，差が 0 であるという前提のもとで特定の確率（5 %など）未満でしか起こりえない大きさのズレであるかどうかを定量的・客観的に決めることができる．

3. 統計的帰無仮説検定の問題と対応

さて，上記のような論理に基づき無事に「有意な差がみられた」という結論を得たなら，学会発表や論文執筆に移りたいところである．しかし，本当にそれだけでよいのだろうか．本節では，統計ユーザーの立場から，考えるべき主要な二つの問題点を挙げ，その対応策について議論する．

3.1 統計的有意性の問題

3.1.1 統計的有意性と実質科学的な「違い」の問題

もう一度，図 1 を見てほしい．直感的に考えて，A と B の違いは「ある」と言ってよいのだろうか．言い方を変えよう．ある子どもがこの図を見て，A と B に違いがあるという結論を導くだろうか．確かに，A の方が分布が右側に偏ってはいる．しかし，その差は直感的にはそれほどでもない，と見ることもできる．

数値を挙げて議論してみよう．B に肯定的評価（4 または 5）を与えている実験参加者が 33% いるということに着目してみる．A の平均値は 3.1 であるということを考えると，A の平均値よりも高く評価している実験参加者が実に 33 %もいることになる．もし，サービスを A に一本化することを考えると，B の 3 割の利用者が無視されることにならないだろうか．それでも，統計的に有意であるから，A が優れているという結論は正しいのだろうか．

検定の第 1 の主要な問題は，何をもって「A が優れている」と考えるのか，その基準設定の問題である．たとえば，「すべての利用者が A の方がよいと判断する」が基準であるとすると，このデータはそれを満たしていない．「平均的に A の方がよい評価である」と考えるのであれば，今回のデータでは満たされているかもしれない．一方で，二つのサービスの平均値を比べてみると僅か 0.2 程度の差である．5 段階評価の 0.2 の差をどう考えればよいだろうか．5 段階評価で違うと言うからには，最低でも「評定値が 1 以上」といった基準を設定すべきという意見もあるだろう．

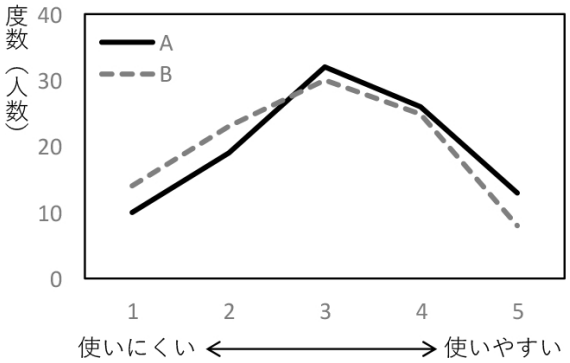


図 4 本稿例（N=200）の度数折れ線グラフ

これらの基準は，検定とは別に定められるべきことである．このような基準設定を，豊田は「実質科学的な要請」と表現している [4]．実質科学的要請とは，どの程度で「差がある」と判断してよいのかについては，検定の外側で，研究者や実務家の責任で決めることを意味する．統計的有意性と実質科学的要請による「違い」は必ずしも一致しない．したがって，検定のみに基づいて「違い」を議論することは，必ずしも適切ではない．

とは言っても，現実的に，基準設定は悩ましい問題である．研究者が決めるとしたら，科学的に正しいかどうかは別として，自分に有利な基準を設定したいと思ってしまうかもしれない．そのような客観性を損なう事態を避けるために，統計的有意性と実質科学的要請による「違い」を同一視して，統計的有意性が一つの基準として使われてきたとも言える．しかし，本来的には，科学的研究は理論的に意味を持つような現象を見出したり，実用に役立つ効果を扱うことを目指すものだろう．そのことを踏まえれば，実質科学的要請に応える基準を見出すべきである．

検定に基準を委ねることが危険であることは，次のサンプルサイズの問題で決定的になる．

3.1.2 サンプルサイズの問題

検定の第 2 の主要な問題は，サンプルサイズ（いわゆる N，実験参加者数，被験者数等のこと）が p 値に多大な影響を与えることである．図 4 をご覧いただきたい．図 1 とそっくりのデータであるが，サンプルサイズを半分の N=200 にしてあり，各サービスの評定を行った人数もちょうど半分にしてある．平均値と標準偏差は同じになる．さて，この結果は有意だろうか．実は，有意ではない（t(198)=1.50, p=.13，両側）．平均値も標準偏差も同じなのに，一方のデータは有意であり，他方のデータは有意ではない．結果が食い違う理由は，p 値の算出にサンプルサイズが影響しているからである．検定では，サンプルサイズが大きいほど p 値が小さくなりやすく，有意になりやすい．

サンプルサイズが大きい方が，母集団の性質をより正確に推定できることは，直感的にも明らかだろう．たとえば，世論調査では調査対象者が多ければ多いほど，結果は正確になる．これは数学的には大数の法則と呼ばれるものである．サイコロを数回振ると，はじめは 3 の目がずいぶん多いといった誤差が生

じるが，振る回数を増やせば，いずれの目の出る確率も 1/6 に近づく．サイコロの場合には，理論的な確率である 1/6 が明らかになっているが，データを取る場合には本当の値はわからない．しかし，試行を増やせば増やすほど，本当の値に近づくことを，大数の法則が保証してくれる．したがって，この考え方から言えば，調査のサンプルサイズは大きければ大きいほど望ましい．

研究の現場を考えると，現実的に集められるサンプルには限界がある．大きなサンプルサイズを確保する努力をするよりも，その努力を新しいアイデアを出すために充てた方が，トータルとしては研究が進むことになるかもしれない．しかし，一方で，お金があって外部にデータ取得を委託できるといった事情があれば，たくさんのサンプルを集めることができる．それぞれの人（研究室）の置かれた条件によって，確保できるサンプルサイズが異なるという現実がある．

ここまでに，サンプルサイズが大きいほど有意な結果を得やすいという事実と，サンプルサイズの大きさは研究者側の都合で変化するという現実を述べた．ここに p 値を用いることの問題がある．つまり，サンプルサイズを大きくしやすい条件が揃った人（研究室）ほど有意な結果を出しやすいということである．これは非常に困った事態である．なぜなら，母平均に差があるかどうかは個人や団体の都合によらない真実のはずである．その真実が，研究者側の都合によって歪められることにもなる．同じ研究をしているにもかかわらず，サンプルサイズを確保するお金があるほど有意な結果を得やすく，結果的に論文も書きやすい，という現実を許してよいのだろうか．

サンプルサイズを増やすと検定の結果が有意になりやすいという問題は，統計的な差と実質科学的に意味のある差の違いの問題とも関係する．扱う現象にもよるが，極端に大きなサンプルサイズを用いることによって実質科学的にはほとんど意味のないような小さな差でも「有意」にすることができてしまう場合がある．このことには，科学的研究やその結果に基づく施策を大きく歪める可能性がある．当該の研究分野にとって実質的には意味がない結果を重要な成果であるかのように見せかける恐れがあるからである．

3.2　問題への対応策

3.2.1　効果量の利用

ここまでに述べた検定の主要な二つの問題は，(1) 統計的有意性と実質科学的要請による「違い」が必ずしも一致しないことと，(2) 研究者側の都合でサンプルサイズが変化するにもかかわらず，サンプルサイズが大きいほど p 値が小さくなりやすいことであった．これらの事実から，p 値のみに頼った判断を避けようとする流れが強まっている．では，具体的にどうすればこれらの問題を克服できるのであろうか．

一つの解決策として近年しばしば用いられるのは，効果量である．本稿例のような 2 群の平均値の比較では，効果量は d で表され，$d=|x_A-x_B|/s$ で定義される．ここで，x_A と x_B は測定された両群の平均値，s は両群のデータから推定された標準偏差であり，2 群の分散をサンプルサイズに応じて比例配分して算出されることが一般的である（詳しくは大久保・岡田 [5] 等の統計書を参考にしてほしい）．

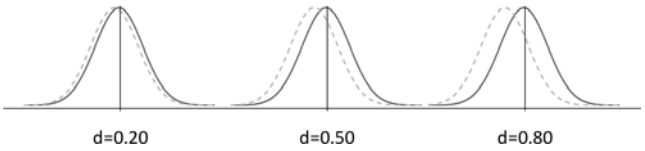


図５　効果量と分布の関係

d の重要な性質は，サンプルサイズに依存しないということである．たとえば，サンプルサイズが N=400 の図 1 のデータでは，d=0.197 である．一方，N=200 の図 4 のデータでも，d=0.197 である．定義上サンプルサイズが含まれていないので，当然である．

d は，確率分布からみた差の大きさについて，客観的に統計的観点から判断することに役立つ．Cohen [6] では，d=0.80 が大きな効果量，d=0.50 が中程度の効果量，d=0.20 が小さな効果量という目安を示している [5, 7]．図 5 に示すように，2 群の確率分布を見てみると，d を直感的に理解しやすい．d=0.20 では，右側の群（実線）よりも左側の群（破線）の方が平均的には小さい値をとるのだが，右側の群の平均値よりも大きい値を取る左側の群をみると，それでも半数弱とずいぶん多い．実際に計算すると，右側の群の平均値よりも大きい値を取る左側の群は約 42% になる．本稿例のデータはこれと同様であり，それほど大きな差ではないと解釈されることが一般的である．一方で，d=0.80 の方は両群の分布がある程度離れている．右側の群の平均値よりも大きい値を取る左側の群は 21% になる．

ただし，どの程度で「差がある」と見なすかは，最終的には実質科学的要請に基づいて判断すべきである．実質科学的要請を無視して一律の基準で判断することは避けるべきである．特定の研究領域において実際に報告された効果量を集めて，分野特有の目安を作る試みもなされている [8]．

効果量については，例に挙げた 2 群の平均値の差以外に，他の様々な分析に対応した指標がある．たとえば，分散分析では η^2 がしばしば用いられる．効果量の詳しい解説は，[6] を参照してほしい．

3.2.2　検定力とサンプルサイズ設計

サンプルサイズが大きいほど有意な結果が得られやすいことは，サンプルサイズが大きいほど微少な差を検出しやすいとも言える．微少な差を検出できることにはメリットもあるが，あまりにも微少な差であると，実質科学的要請に照らし合わせて意味が無いかもしれない．そこで，あらかじめ「母集団で想定される差」を決めておき，その差を検出するのに適切なサンプルサイズを決定しようという発想がある．これが，検定力分析に基づくサンプルサイズ設計である．近年，心理学の分野で積極的に取り入れられ始めている [9]．

検定力分析の説明には，検定におけるエラーの概念が必要であるので簡単に解説する．先に述べたように，検定は，母平均に差がないという前提のもとで，手元にあるサンプルが得られる確率を計算する．通常，この確率が 5% 未満であれば，母平均に差がないというはじめの前提（帰無仮説）が誤りであると

判断する．しかし，あくまでも確率に基づく判定であるので，本当は差がなくても誤って帰無仮説を棄却することもある．これはタイプⅠエラーと呼ばれる．タイプⅠエラーの確率は有意水準 α で表され，通常 α =5% に設定される．

上記は母平均に差がないという前提のもとで計算をはじめたが，この前提を「母平均に差がある」に変更してみよう．母平均に差がある場合には，差を具体的に固定しないと計算ができない．そこで，本稿例を考えて，平均値の差を 0.5 とする．5 段階評価の中で，0.5 程度の平均値の差があれば，違いがあると考えよう，という想定である．また，詳細への言及は避けるが，計算上，母集団の標準偏差も合わせて想定しておく必要があるので，これは 1.2 とする．この前提のもとで，適当なサンプルサイズを決めれば，そのサンプルサイズに応じて，有意な結果が得られる確率が求められる．手順としては，図 3 に示した方法と同じである．上記の想定でサンプリングをくり返すと，平均値の差の分布が得られる．

仮に A と B の各群について N=100 として，有意であると判定される確率を求めると，83% である．この確率が検定力と呼ばれるものであり，しばしば 1-β で表現される．検定力とは，母平均に差があるという前提のもとで，正しく差があるという結果を返す確率である．もちろん，確率であるから誤りの可能性もある．この誤りはタイプⅡエラーと呼ばれ，β で表される．この場合は β =17% である．

上記ではサンプルサイズを固定して検定力を求めたが，逆に一定の検定力を得るためのサンプルサイズを求めることもできる．たとえば，検定力として 1-β =90% はほしいとしよう．N の関数としたときの検定力は図 6 の真ん中の線のようになる．両群 N=122 のときにほぼ 1-β =90% となる．したがって，両群 N=122 になるように調査を計画する．これが検定力分析に基づくサンプルサイズ設計である．

このとき，余力があるからと言って，N=122 を大きく超えるサンプルを集めないことが肝心である．N を増やしてしまうと，必要以上に微少な差を検出してしまう可能性が高まるからである．たとえば，図 6 には母平均の差が 0.3 である場合のサンプルサイズと検定力の関係も載せた．サンプルサイズが 300 近くになると，母平均の差が 0.3 であっても，80% 以上の確率で差を検出することになる．ここでは，必要以上に小さな差を検出しないように，あらかじめ差の大きさを想定しておき，その差をおよそ検出できるようにサンプルサイズを決めようとしている．こうすることで，必要以上に小さい差を検出することを避けるのである．とは言っても，実務上の難点は多い．一つは，検出したい差をどのように決定するかという問題がある．最終的には実質科学的要請に基づくことになるが，先にも述べたように，決定はなかなか困難である．また，別の問題として，サンプルサイズが大きいほど推定としては正確であるのに，わざわざサンプルサイズを小さくすることを正当化する側面がある．可能な範囲で大きいサンプルサイズを確保した方が，より正確な結果が得られるはずなので，基本的にサンプルサイズが大きいことは統計的には好ましい．しかし，参加対象に過度の負担をかけない（不必要にサンプルを増やさない）で研究を実施するという点で研究倫理的，また，経済的には好ましいという見方もある．いずれにしても，実際に収集したサンプルサイズが大きい場合には，次に述べる信頼区間を使用するといったように，検定以外の方法を合わせて吟味した方がよいだろう．

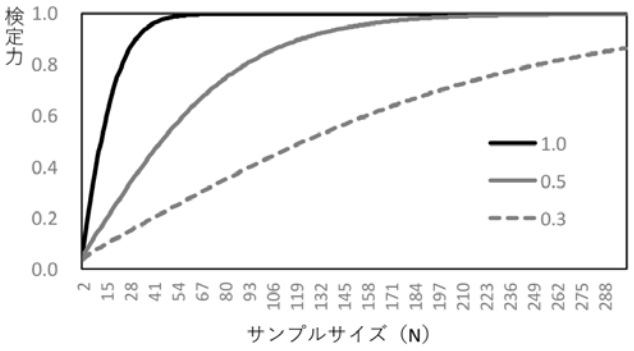


図６　検定力

3.2.3　信頼区間の利用

サンプルサイズが大きい方が，推定の精度はよい．しかし，図 1 と 4 のデータを比べると，平均値の差は同じ 0.23 であるし，効果量も同じ 0.197 である．本当はサンプルサイズが大きい方が精度がよいはずなのに，平均値の差や効果量をピンポイントで表現するこれらの指標は，推定精度を表現していない．では，サンプルサイズが大きいことを活かした別の表現方法はないだろうか．その一つが信頼区間である．

本稿例において，2 群の平均値の差の信頼区間を計算すると，95% の信頼度で [0.001, 0.459] である．このように，ピンポイントではなく，幅を持たせて差を吟味するのが，区間推定の考え方である．区間推定はさまざまな対象に対して可能であり，たとえば A の平均値は 3.13 であるが，この平均値を区間推定すると 95% の信頼度で [2.97, 3.29] となる．

信頼区間が有用な理由は，サンプルサイズを反映するからである．サンプルが半減した各群 N=100 の場合，2 群の平均値の差の 95 %信頼区間は [-0.096, 0.556] である．また，A の平均値は [2.90, 3.36] である．2 群の平均値の差も，A の平均値も，サンプルサイズが大きい場合の方が区間の幅が狭いことが重要である．すなわち，区間の幅により推定の精度を表現することができている．APA（American Psychological Association；アメリカ心理学会）[10] の出版マニュアルによれば，効果量とともに信頼区間を表記することが勧められている．このように，サンプルサイズを考慮して，かつ平均値の実質的な差を吟味する手段として，信頼区間を利用できる．

ただし，信頼区間の解釈はしばしば誤解しやすいので，注意が必要である．よくある誤解は「信頼区間の中に真の値が存在している確率が 95% である」という解釈である．正確には「真の値が存在している確率は 0 か 1 のどちらか（ただし，どちらかはわからない）」という解釈が正しい．なぜなら，真の値は変動するのではなく，信頼区間を計算する前にあらかじめ決められていると考えるからである．これは，ここまでに述べたように，検定では理論上母平均の差をあらかじめ決めておくことと同じである．

信頼区間は，図 3 において説明したように，サンプリングの繰り返しによって定義される．1 回サンプリングするごとに，平均値の差の信頼区間を「一定の計算手続き」に従って計算する．これを何度もくり返す．もし図 3 に示したように，母平均の差が 0 であるとしても，サンプルは確率的に変動するため，信頼区間は 0 を含むこともあれば，0 を含まないこともある．このようにくり返しサンプリングをした場合に，95% の確率で 0 を含むことが保証されるように定義されるのが，信頼区間である．区間の幅を広く取るようにすれば 0 が含まれる（二つの平均に差がない）確率が高くなるし，狭く取るようにすれば 0 が含まれない（二つの平均に差がある）確率が高くなる．その確率がちょうど 95% になるように決めたものが，95% 信頼区間である．

信頼区間に対する誤解は，検定も同様だが，理論的には母集団の性質を仮定して議論しているのに，現実には母集団の性質がわからないから調査をしているという矛盾から生じている．現実の調査では，事前に母集団の性質を仮定することは不可能である．調査を行う場合には，母平均はわからないし（わからないから調査をするので当然），手元のサンプルは 1 回分しかない．近年，この問題を克服できる手段の一つとして，ベイズ統計が広く使われるようになってきた．これについては，第 5 節で述べる．

3.2.4　p 値の解釈に対する二つの立場

ここまで，p 値の問題点を述べてきたが，p 値を積極的に解釈に活かそうという立場もある．p 値の解釈に対しては，次の二つの立場がある [11].

第 1 に，ネイマン・ピアソン流の立場である．5% が強調される背景にもなっている．この立場では，仮説が正しいか，誤りかの 2 者択一の考え方をとる．p 値が 5% を下回ることが重要であって，下回った場合に p 値がどの程度であったのかという点は考慮されない．有意水準は事前に α =.05 といった形で決められるべきであり，事後の変更は許されない．検定力分析に基づくサンプルサイズ設計の考え方は，この立場に基づいている．

第 2 に，フィッシャー流の立場である．この立場では，p 値は差があるかどうかを評価する一つの指標であって，5% という基準に関わらず，p 値を積極的に解釈する．p 値は低ければ低いほど，証拠としての強さがあると考える．サンプルサイズによって p 値が影響されることは問題ではあるが，サンプルサイズが大きいほど推定が正確になるのだから，p 値が小さいほど証拠として認めようという立場は理解できる．

心理学の分野では，従来はネイマン・ピアソンの立場から 5% 水準を基準としてきたが，現代では 5% 水準を使いつつ，フィッシャーに近い立場が取り入れられている．APA[10] では，論文に記述する p 値を正確に報告するように勧めている．すなわち，p<.05 といった書き方ではなく，p=.023 といった書き方が推奨される．これは，p 値が小さいほど証拠として強いと認めようという立場の現れであると言える．現実的には，これまでにコンセンサスが得られている 5% を基準としつつも，5% で 2 者択一的な解釈を行うのではなく，より柔軟に p 値を解釈する流れになっている．

4. 統計ユーザーとして心がけること

ここまでに，検定における主要な二つの問題点とその対応策を述べてきた．しかし，統計的な理論や技術とは別に，統計ユーザーの立場からもいくつかできることがある．本節では，その中からさらに 2 点を取り上げて，問題点と対応策を考える．

4.1　統計的帰無仮説検定の前にやるべきこと

アカデミックな領域でユーザーの立場から実用的な統計学を学ぶと，検定にどうしても注目しがちである．それは，学会発表や論文の中で検定が頻繁に使われていることからも理解できる．しかし，平均値や標準偏差などの基本統計量の確認と，ヒストグラムや散布図などによるデータの可視化こそが，検定に先立って必要な，極めて重要な工程である．極端な例では，検定の結果のみ記述され，肝心の平均値が書かれていない研究発表を見ることがある．検定は重要な手法ではあるが，それが唯一の手法ではないし，問題も多いことをここまでに述べてきた．検定を行う前に，まずは目の前のローデータと向き合い，基本的な性質を探ることが必要である．

本稿例のようなデータでは，平均値と標準偏差をはじめに確認したい．また，図 1 のように可視化することで，データの特徴をつかむこともできる．たとえば，この段階で評定値が 4 と 5 に極端に偏っていたとしよう．このまま検定を行ったとしても，偏ったデータでは多くの検定の前提となる正規性が仮定できないし，バラツキが過小評価されてしまっていると考えられる．もちろん，最終的には研究目的によるが，そのような指標を外して分析したり，適切な指標で再評価したりした方がよいかもしれない．二つ以上の指標同士の関係性を検討したいのであれば，クロス集計表にまとめたり，相関係数や散布図を確認したりすることが必要であろう．検定に先立って，基本的なデータの性質を確認することを忘れないようにしたい．

4.2　2 分法的な考え方から解放されること

あらかじめ 5% という基準を設け，機械的に効果があった，なかったと判定することは楽である．いままで，我々を含めた統計ユーザーは，この基準を当たり前のように受け入れてきた．ここまでに述べてきた問題が議論され，この基準が大きく揺らいでいるのが現代である．このような基準が受け入れられてきた背景には，我々が，仮説が正しいか，誤りかという 2 分法的な解釈を使ってきたことがある．この背景にはいくつかの理由がある．

一つは，研究する人間の思考の問題である．2 分法的な解釈ができれば理論的に簡潔であるし，選択肢が二つしかなければ，選択自体も楽である．中間的に決着がつかないまま時間を浪費するよりも，決着をつけて次の段階に進む方が，気分的にもすっきりするし，生産性の面から見ても合理的なのかもしれない．

また，研究制度の問題もある．査読には通るか，通らないかの 2 者択一であって，中間的な解決は今のところない．そのよ

うな決着をつけるためには何らかの「ルール」が必要になる．研究者の評価にも直結するし，若手研究者は就職（生活）がかかっているから，ルールの公平性も求められる．このような要請から，一見公平に見え（本当は問題も多いが），決着をつけてくれる検定を共通ルールとするのは合理的なのかもしれない．

しかし，これらは，我々が研究する本当の目的である，真実を明らかにすることとは本来関係がない事情である．それまでの研究の積み重ねと新たに得られたデータから，新しい事実を明らかにすることが研究の本質である．それは，簡単に合っている，間違っていると判断できるものではない．特に制度の問題は難しいが，2 分法による判断はデータを扱う本来の目的からは乖離があるということを，研究者が認識しておく必要はあるだろう．

5. 近年の動向

ここまでに，検定に関わる問題点と対応策について述べてきた．上記は比較的以前から議論が続けられてきたことであるが，ここでは最近 10 年程度の間に活発になった，検定を含めた統計的分析の変化について紹介する．

5.1 メタ分析

メタ分析とは，いくつかの類似の研究データを統合し，大きなサンプルサイズのもとで分析をする手法である．母集団の性質を正確に推定するには，1 回の調査だけでは限界がある．メタ分析は，この問題を解決する一つの手法であり，医学系の分野を中心に，近年盛んにメタ分析を用いた論文が刊行されている．たとえば，ハッティ [12] では，少人数学級や宿題に学習効果があるのかといった教育における問題について，多数の研究のデータを統合し，その効果量を推定している．類似の研究データが多数集まることが条件ではあるが，データから何らかの学術的な判断を下す場合には，1 回の調査のみではなく，メタ分析的なアプローチが重視されるようになってきている．また，検定力分析に基づくサンプルサイズ設計を行うためにも，メタ分析の結果は役立つ．メタ分析で推定された効果量を検定力分析に用いるのである．加えて，これらの効果量は個別の研究の実質科学的な意義を解釈する際にも役立つ．

5.2 再現性

研究結果の再現性の問題が心理学を中心に話題になっている．科学的研究は再現性が求められるにもかかわらず，多くの実験が追試に成功しないという指摘が相次いだ．日本の心理学関係で，もっとも著名なレビュー雑誌である『心理学評論』でも再現性問題の特集が組まれるほどの関心の高さである [1]．再現性問題の原因にも，検定の問題が絡んでいる．

たとえば，似たような実験を数回くり返していれば，母集団では差がなくても有意な結果が得られることはある．典型的なタイプⅠエラーである．別の言い方をすれば，何度も実験を繰り返せば，論文として発表できるデータが得られてしまう．このように有意な結果を人為的に生み出すことを，p-hacking と呼ぶ（たとえば，[13]）．他にも，たくさんの指標をとって有意であったもののみ報告する，有意になった人数の時点で調査や実験を打ち切るといった p-hacking がある．これらは，悪意が無くても行ってしまうことはあるだろうし，問題を認識しつつもデータ取得の都合上行ってしまうこともあるだろう．

この問題を解決するために，二つの方向性がある．一つは，データをすべて正直に報告することである．すなわち，多数回の実験を行ったのであればそのように，多数の指標をとったのであればそのように報告することである．ただし，論文にそれらの詳細を記述することは困難であるし，どこまで報告してよいのか現時点ではコンセンサスがない．さらに，ネガティブデータがあることで有効なデータの論文が掲載されないということにでもなれば，研究者としての評価に直結してしまうので，すべてを正確に報告することをためらう気持ちは理解できる．

この問題を解決できる，もう一つの方向性として，実験・調査の事前登録制度がある．日本では，心理学の代表的な雑誌のひとつである『パーソナリティ研究』がこの方式を採用することになっている [14]．これは，実際に実験や調査を行う前にその手続き（予定しているサンプルサイズや分析方法を含む）について査読を受けるものである．受理されたならば，事前登録された実験・調査を忠実に行う．実施された研究は，結果が望ましいものでなくても，論文として掲載される．このような体制をとることで，p-hacking を防ぎ，同時に p-hacking を行なおうとする動機づけもなくなることが期待される．心理学では科学的探究の側面が重視されるので，開発的な側面が重視される工学研究とは異なるかもしれないが，このような体制を築くことで，研究者の利益も保証されるし，データの信用性も上がると思われる．今後注目すべき取り組みである．

5.3 データの特徴を考慮した分析

たとえば，100 の中学校を対象に，各学校の 3 年生から 1 クラスを抽出し，勉強に関する意識調査をしたとしよう．1 クラスは便宜上 30 名としておく．典型的な分析手法の一つは，30 名 ×100 校 = 3000 名をサンプルとして分析することである．しかし，このデータで検定を行う場合，一つの問題がある．それは，この 3000 名は独立サンプルではない，すなわち 1 クラス 30 名は同じクラスであることによって意識調査の結果に何らかの系統的な関係があると推測され，検定の前提であるランダムサンプリングが満たされていない．実際，このようなデータを独立サンプルとして分析するとタイプⅠエラーを増大させてしまう [15]．そのため現代では，このようなデータには「マルチレベルモデル」と呼ばれる，同じクラスの 30 名に関係があることをあらかじめ計算に組み込んだモデルを使うことが勧められている．このように，データの特徴を考慮した分析が行われるようになってきた [16]．

従来は，t 検定や分散分析，重回帰分析といったような，あらかじめ決められた，限られた数の分析手法に合うよう，「データの細かい前提」を捨てて，データを分析手法に（無理矢理）合わせていた側面がある．一方で，近年では統計ユーザーが利用できる分析手法が増加し，上記のマルチレベルモデルのように，データの特徴に合った分析手法を実用的に選べるようになってきた．さらに柔軟に，データの細かい前提までもモデル

に組み込んでモデル化（モデリング）し，分析することが可能になってきた．このような発想は，次に述べるベイズモデリングとの相性がよい．

5.4　ベイズ統計

ベイズ統計は，ここまでに述べた体系とは異なる，ベイズ流の統計学を応用したものである．これまでの検定を中心とした統計分析手法を置き換えるものとして紹介されることもある [4]．ここでは，本稿例をベイズ統計で分析してみよう．検定では，2 群の母平均が等しいこと（あるいは等しくないこと）を前提に議論を進めるが，ベイズ統計では母平均は未知として，今得られているデータから母平均を推測するという立場を取る．詳細は参考文献に譲るが，豊田 [4] の方法に沿って平均値の差を推測すると，図 7 のような平均の差のヒストグラムが得られる．これは母平均値の差を推定した確率分布である．ここから母平均の差は 0.2 あたりの値である確率が高いことがわかる．実際，このヒストグラムの平均値は 0.230 であった．また，両端の 5% を除き，95% が含まれる範囲をとると，[0.004,0.459] であった．これは 95% 確信区間と呼ばれる．また，各群 N=100 のデータでは，95% 確信区間は [-0.100, 0.561] となった．

信頼区間と確信区間は異なる概念である．確信区間は「区間の中に真の値が存在している確率が 95% である」を意味していて，直感的に理解しやすい．直感に合う理由は，得られたデータから母集団の性質を推測するという，研究者が置かれた状況に合致しているからである．この意味で，ベイズ統計は自然な方法として理解できる．

また，ベイズ統計は，先に述べたモデリングとの相性が良いことも知られている．ベイズ統計が普及した背景に，MCMC 法（マルコフ連鎖モンテカルロ法）と呼ばれる数値計算手法が発展したことがある．これにより，さまざまな柔軟なモデルを組んでも，数値計算ができるようになった．現在では検定が主流となっている分野においても，今後主な分析手法がベイズ統計に置き換えられていく可能性もある．

5.5　ビッグデータの分析

ディープラーニングに代表される最近の AI は，多くのデータから学習をすることによって発展してきた．また，ICT（情報通信技術）の発展で，さまざまなデータを取得することが容易になっている．このような膨大なデータはビッグデータと呼ばれる．ビッグデータに対しては，検定は無意味であることが多い．なぜなら，サンプルサイズの問題で述べたように，ビッグデータは莫大なサンプルサイズのため，ほとんど意味が無い微少な差を検出しやすくなるからである．したがって，ビッグデータの分析を考えた場合には，検定の使用を根本的に見直す必要がある．

6. 読書案内

本稿と関連する話題についてさらに情報を得るために，日本語で読める基本的な文献を紹介する．効果量とサンプルサイズ設計については，[17] が参考になる．心理学分野における検定の問題と改善策についてまとめられている．ベイズ統計については，従来の検定の置き換えとして，[4] が勧められる．読み進めながら実際に手を動かして分析してみることができる．モデリングについては，ベイズ統計の観点から [18] が勧められる．モデリングの具体例が豊富にある．最後に，p 値の理論について詳しく知りたいという場合には，医学分野を対象としたものであるが，[11] を勧めたい．いずれも分野はさまざまであるが，特定の分野に限られた問題は少なく，論点はいずれの分野も共通であることが多い．

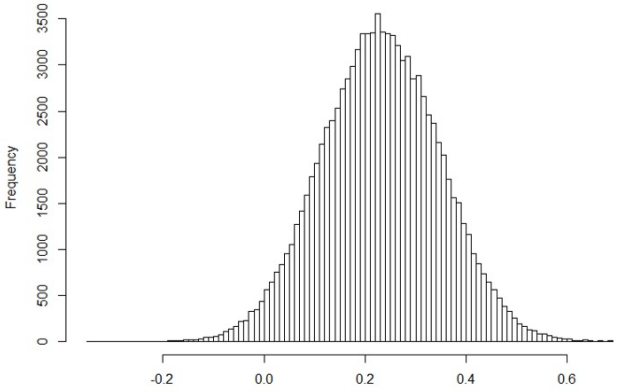


図 7　ベイズ統計による分析結果

7. おわりに

検定の問題点とその改善策について述べてきた．検定には長い間使われ続けてきた歴史があり，研究者のコンセンサスはそう簡単には変わらないだろう．しかし，問題点がここまで明らかになったいま，このまま前例を踏襲し続けることはないだろう．少なくとも，検定「だけ」に頼った分析は無くなるだろう．

筆者を含め統計ユーザーとしては，統計学の勉強に割ける時間が限られている中，統計利用の変革期に遭遇することになり，なかなか困難な時代になった．いくつか今後の見通しを述べたが，それらが現実のものとなるかは未来にならないとわからない．一つのポイントとしては，検定に問題点があることを共通の認識としつつ，自身の研究領域の動向を注視することだと思う．おそらく，多くの先行者が試行錯誤するはずで，一般的な統計ユーザーとしては，ある程度コンセンサスが得られてから本格的に参入してもよいのではないかと思う．もちろん，当該領域の先行者として活躍することも一つの道であると思う．

謝辞

本稿の執筆にあたり，JSPS 科研費 JP16H03073, JP17K18620 の助成を受けた．

参　考　文　献

[1] 友永雅己, 三浦麻子, 針生悦子: “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―特集号の刊行に寄せて―,” 心理学評論, Vol.59, No.1, pp.1-2, 2016.

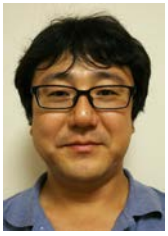
[2] 三浦麻子, 岡田謙介, 清水裕士: “統計革命：Make statistics great again―特集号の刊行にあたって―,” 心理学評論, Vol.61, No.1, pp. 1-2, 2018.

[3] 田中敏, 中野博幸: R&STAR データ解析入門, 新曜社, 2013.
[4] 豊田秀樹: 新訂 心理統計法―有意性検定からの脱却―, 放送大学教育振興会, 2017.
[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, 勁草書房, 2012.
[6] J. Cohen: *Statistical Power Analysis for the Behavioral Sciences*, Revised Edition, Academic Press, 1977.
[7] 森敏昭, 吉田寿夫: 心理学のためのデータ解析テクニカルブック, 北大路書房, 1990.
[8] P. E. Morris & C. O. Fritz: "Effect sizes in memory research," *Memory*, Vol.21, No.7, pp. 832-842, 2013.
[9] 井関龍太: "実験心理学における例数設計の周辺," 基礎心理学研究, 印刷中, 2019.
[10] APA (American Psychological Association): *Publication Manual of the American Psychological Association*, Sixth Edition, American Psychological Association, 2010.
[11] 柳川堯: P 値 その正しい理解と適用, 近代科学社, 2018.
[12] ジョン・ハッティ: 教育の効果 メタ分析による学力に影響を与える要因の可視化, 図書文化, 2018.
[13] 藤島喜嗣, 樋口匡貴: "社会心理学における "p-hacking" の実践例," 心理学評論, Vol.59, No.1, pp. 84-97, 2016.
[14] 加藤司: "『パーソナリティ研究』の新たな挑戦―追試研究と事前登録研究の掲載について," パーソナリティ研究, Vol.27, No.2, pp. 99-124, 2018.
[15] 村山航: "刺激の効果を侮るなかれ―ランダム刺激効果を含んだ線形混合モデルの重要性と落とし穴," 基礎心理学研究, Vol.36, No.2, pp. 236-242, 2018.
[16] 熊谷龍一, 荘島宏二郎: 教育心理学のための統計学, 誠信書房, 2015.
[17] 村井潤一郎, 橋本貴充: "心理学のためのサンプルサイズ設計入門," 講談社, 2017
[18] マイケル・D・リー, エリック-ジャン・ワーゲンメイカーズ: 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, 2017.

（2019 年 4 月 8 日　受付）

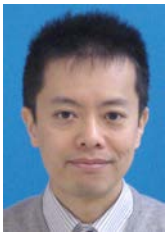
［問い合わせ先］
〒380-8544　長野県長野市西長野 6-ロ
信州大学
島田　英昭
E-mail: hshimada@shinshu-u.ac.jp

著 者 紹 介



島田（しまだ）　英昭（ひであき）［非会員］

2004 年筑波大学大学院心理学研究科修了．博士（心理学）．現在，信州大学学術研究院教育学系教授．日本心理学会，日本教育心理学会，日本認知科学会，日本教育工学会，日本デジタル教科書学会などの会員．
http://shimadahideaki.jp



井関（いせき）　龍太（りゅうた）［非会員］

2005 年筑波大学大学院心理学研究科修了．博士（心理学）．現在，大正大学心理社会学部講師．日本心理学会，日本教育心理学会，日本認知科学会，日本認知心理学会，日本読書学会などの会員．