帰納的推論ツールとしての統計的仮説検定

— 有意性検定論争と統計改革 —

Statistical Hypothesis Testing as a Tool for Inductive Inference

— the significance test controversy and the statistical reform —

土居淳子

1. はじめに

統計的仮説検定は，主観によらない「客観的」判断をくだすための「科学的方法」として心理学をはじめとする社会科学の広範な領域で盛んに用いられている。そのため，社会科学系の大学教育カリキュラムには統計的仮説検定を主とした推測統計法に関する授業が組み込まれ，また，膨大なテキストや解説書が出版されている。

ところが，社会科学領域の学生・研究者向けに書かれたテキストにおいて「統計的仮説検定」として紹介されている手法は，フィッシャー（Fisher, R.A.）が1920年代に展開した有意性検定論とも，また，フィッシャーの理論を土台としてネイマン（Neyman, J.）とピアソン（Pearson, E.S.）が1930年代に展開した統計的仮説決定理論とも異なる。1940年代，フィッシャーとネイマン=ピアソンがそれぞれの仮説検定論の是非をめぐって激しい論争を繰り返し[1]，方法論としての合意が得られないまま統計ユーザー向けに折衷案として登場したのが現在の統計的仮説検定，いわゆる「(帰無仮説)有意性検定」である(Gigerenzer&Murray, 1987)。

互いに相容れないフィッシャー理論とネイマン=ピアソン理論を，無理やり合体させることで誕生したともいえるこの有意性検定は，その誕生直後から現在に至るまで，様々な厳しい批判を受けてきた（レビューとして，Kline, 2004; Fidler, 2005; 橘 1986)。有意性検定

[1] フィッシャーとネイマン=ピアソンの論争は現在に至るまで合意に至っていない。今日の数理統計学界は，ネイマン=ピアソン学派，フィッシャー学派，ベイズ学派の3学派に大別され，各理論の立場に依拠した方法が未整理のまま割拠している(芝村，2003)

の認否をめぐる，1970年以前の15年間に現れた論争の論文集がモリソンとヘンケルによって編纂され（Morrison, D.E.&Henkel, R.E., 1970），その邦訳が10年遅れで内海ら（1980）によって上梓されている。Rozeboom（1960）は，この論文集に収録されている論文“The fallacy of the null hypothesis significance test（帰無仮説の有意性検定の誤謬）”において次のように述べている。

> 私はこの論文で，少なくとも心理学者にたいしては宗教的な信念の地位を得るにいたった推論手続きの教義を検討しようと思う。吟味されるべき教義は，「帰無仮説の有意性検定」への信仰である。…… 実験科学雑誌や応用統計学の教科書でこの方法はおそるべき高い地位をえてはいるものの，この方法の基礎には合理的推論の性質にたいする根本的な誤解があり，科学的研究の目的に適していることはめったにない，ということである。（Morrison&Henkel，1970，訳書 p.204）

また，Bakan（1967）は，“The test of significance in psychological research（心理学研究における有意性検定）”において，次のように述べる。

> この論文で語られることはほとんど独創的なものではない。それは，ある意味では「周知」のことがらである。それを「大声で」言うことは，いわば王様がほんとうは裸であると指摘した子どもの役割をひきうけることにあたる。（Morrison&Henkel, 1970, 訳書 p.223）

しかし，これらの有意性検定批判は，橘（1986）の言葉を借りれば，「徹底的に無視され続けてきたために，多くの研究者には気づかれずに現在に至っている（p.16）」。橘（1986）もまた，有意性検定の適用が儀式化していることに対して

> 有意性検定は「さぼり」の論理で，単なる便利な，それゆえに安易な判断を行うための道具でしかない。便利であることは妥当であることとは別のことである。ところがほとんどの“科学的”研究で（便利であるから）これを使っているために，これを使わないで結論を出すことは非科学的であるというように勘違いしてしまったのである。（橘，1986，p.97）

と苦言を呈している[2]。

しかし，近年になって変化が生じつつある。Cohen（1994）が“The earth is round（$p < .05$）”

橘（1986）は，一般の統計ユーザーが理解しうる平易な説明を用いて，有意性検定のかかえる根本的な問題とそこから生じる誤用と弊害を指摘している。

と題する論文において有意性検定に対する辛辣な批判を行い，有意性検定に代わる統計手法として効果量の区間推定を提案したことが直接の契機となり，アメリカ心理学会（American Psychological Association，以下 APA と略す）は 1996 年に推測統計に関する専門委員会を設置した。1999 年には心理学研究において推定・検定を使う場合のガイドラインが委員会提案として発表され，有意性検定を行う際には検出力を考慮して標本の大きさを事前決定すること，有意確率（p 値）だけでなく効果量を明記すること，信頼区間を図示することなどが，より適切な推測統計の利用方法として推奨された（Wilkinson ら，1999）。

本稿では，まず 2 節でフィッシャーの有意性検定論とネイマン=ピアソンの統計的仮説決定理論を簡潔に紹介し，現在の統計的仮説検定，すなわち有意性検定を両理論の折衷法と位置付ける。次に 3 節では，折衷案として誕生した有意性検定が本質的に内包する問題に起因する p 値の解釈をめぐる様々な誤解と，第 2 種の過誤確率の制御（検出力）の問題について述べる。また，このような弱点を自覚せずに有意性検定を「科学的方法」として用いることが，科学的推論に悪影響を及ぼした事例を紹介する。最後に 4 節で，APA を中心に進められつつある統計改革の内容とその浸透状況を手短に紹介し，また，日本における統計改革についても若干言及する。

なお，本稿で述べることの多くは，すでに他の研究者によって半世紀にわたって何度も繰り返し述べられていることである。したがって，学術論文としてのオリジナリティはここにはない[3]。また，本稿で述べることは，多くの社会科学研究者にとって周知の事実であるかもしれない。しかし，入門的な統計テキストではほとんど触れられないため，有意性検定にまつわる論争や批判を知らないまま，難解で分かりにくくなかなか腑に落ちない有意性検定の論理構造に頭を悩ませている学生・研究者も少なくないと思われる。筆者自身，一般的なテキストの仮説検定に関する記述を読むたび，論理的に整合していないと思われる結果解釈に出会うことが多く，もやもやとした思いを抱えることが多い。本稿は，そのような統計ユーザーの「もやもや」を解消する一助になればと思う。

[3]同様の言明を，少なくとも 1967 年に Bakan が，1996 年に Cohen が行っており，この言明自体にもオリジナリティは全くない。

2. フィッシャーの有意性検定論とネイマン=ピアソンの統計的仮説決定理論

フィッシャーは農事試験場の統計研究員であった1920年代に有意性検定論を展開している。芝村（2004）によれば，フィッシャーが有意性検定の手続きの形式化を進めたのは，肥料を購入する人々への説得をより容易にするためであった。

図1は，この検定論の手続きを示したものである。この検定論では，まず帰無仮説が真であると仮定する（①）。帰無仮説には研究者が反証しようとする仮説を設定するが，「帰無（null）」とは「特徴のない」や「存在しない（ゼロの）」という意味であり，帰無仮説となる仮説が効果や差異の不在を意味する仮説であることを含意している（細谷, 2002, p.94）。次に，その帰無仮説のもとで標本統計量の実現値（④）が得られる確率を計算する。そして，その確率と5%や1%といった有意水準とを比較することで標本統計量の実現値の有意性を査定し（⑤），有意であると判定されれば帰無仮説を棄却し（⑥），有意でない場合は帰無仮説の真偽についての判断を保留する（⑦）。

フィッシャーの有意性検定には，対立仮説という概念はない。帰無仮説が偽であることを前提とし，それがどの程度の確からしさで示唆されるかを手元のデータを用いて測る「帰納的推論」のための方法論である。有意水準をデータの有意性つまり帰無仮説からの乖離の度合いを示すものとみなし，いかなる大きさの有意水準で棄却できたのかが重要となる。有意水準5%よりも有意水準1%の方が，帰無仮説が正しくないであろうと考えられる度合いが強いという解釈に基づくため，フィッシャーの有意水準は確信確率[4]（fiducial probability）とも呼ばれる

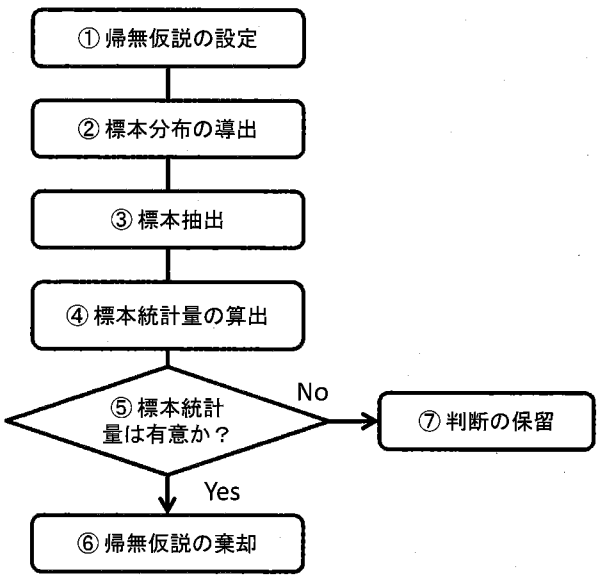

図1 フィッシャーの有意性検定論（木村, 1992, 一部改変）

[4] 信頼確率と呼ぶこともある。

（渋谷＆竹内，1962）。フィッシャーは，自身が強く批判したベイズ統計の「逆確率」と類似の働きを確信確率（有意水準）に求めようとしたのであるが，このような解釈に論理的な整合性はなく一種の自己矛盾を抱えたものであった。

一方，ネイマン=ピアソンの統計的仮説決定理論は，フィッシャーの方法を再構成して1930年代に展開された（図2）。規格化された工業製品の大量生産を背景とする統計的品質管理を主たる応用領域としたため，科学的知見を導きだすための推論ツールというよりはむしろ，毎回の検定ごとに品質の合格・不合格をいかに合理的に判断し行動すべきかを示す意思決定のためのツールである。

ネイマン=ピアソンの理論では，統計的仮説として検定仮説と対立仮説という2種類の仮説を設定する（①）。帰無仮説に相当する検定仮説が採択・棄却の対象となり，検定仮説が棄却されれば対立仮説を採択する。フィッシャーが同一の検定を繰り返すことを前提としないのに対して，ネイマン=ピアソン理論は同じ検定を何度も繰り返すことを前提としている。その繰り返しの中で，単一の検定ごとに検定仮説の採択または棄却という二者択一方式の判断（決定）を行い，その判断には一定の確率で間違いが含まれることを許容するかわりに，検定を何度も繰り返した時のトータルの判断ミスの割合を一定以下に抑えるという理論である。

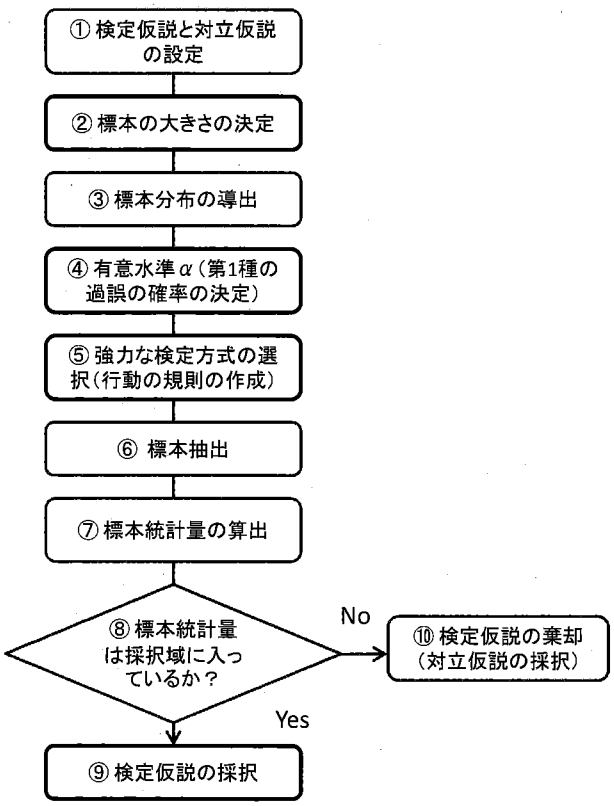

図2　ネイマン=ピアソンの仮説決定理論
（木村，1992，一部改変）

ネイマン=ピアソン理論では，フィッシャーでは主観的・原始的な確からしさの尺度でしかなかった有意水準を，検定仮説が真であるのにこれを棄却してしまう誤り（第1種の過誤）の発生頻度として明確かつ客観的に定義し，さらに，検定仮説が偽（対立仮説が真）であるのにこ

表1　2つの仮説検定論（木村，1992より一部改変）

	ネイマン=ピアソン	フィッシャー
統計的仮説の数	2つ（検定仮説と対立仮説）	単一（帰無仮説）
標本の大きさ	事前決定を要す	事前決定不要
α	第1種の過誤確率	有意水準
β	第2種の過誤確率	なし
$1-\beta$	検定力	なし
意味	相対頻度	確信確率
αの決定	標本抽出以前	標本抽出以後
標本抽出	反復抽出	1回
判定回数	反復	1回
判定内容	検定仮説の採択・棄却	帰無仮説の棄却・判断留保

れを採択してしまう誤りを第2種の過誤として概念化している。そして，標本の大きさと有意水準を事前決定した後（②，④），第2種の過誤確率が最も少なくなるように検定方式（行動の規則），つまり棄却域をどのように設定するかを決め（⑤），第2種の過誤確率がいくらになるかを算出する。

この手続きを用いると，第2種の過誤確率を一定値以下に抑えるのに必要な標本の大きさを事前に知ることができるため，合理的に二者択一的判断（検定仮説の採択⑨または対立仮説の採択⑩）を行うことができる。

ネイマン=ピアソン理論は，フィッシャーの有意性検定において不明確であった点を解決し，数学的に厳密に定式化したものであるとみなすことも可能であるが，一方で，フィッシャーの「帰納的推論」からネイマン=ピアソンの「帰納的行動（決定）」へと目的が大きく変貌している。また，その手続きにおいても多くの相違がある。ネイマン=ピアソンの，検定仮説を機械的に「採択」または「棄却」にふるい分ける方法は，たしかに品質管理などの応用分野では有用かつ合理的な意思決定ツールであるが，科学的推論ツールとして帰無仮説と手元にある標本との乖離の度合い（有意性）を正確に査定しようとするフィッシャーの観点とは相容れないものである。ネイマン=ピアソンとフィッシャーの仮説検定をめぐる対立の要点は表1の通りである（木村，1992）。

3. 社会科学における「推論革命」とハイブリッド化された仮説検定

前述のように，フィッシャーとネイマン=ピアソンはそれぞれの立場の違いから仮説検定理論をめぐって深刻な論争を繰り広げたが，その一方で，統計的仮説検定は研究データの評価に客観性をもたらす非常に魅力的な手法として生物学や社会科学領域に浸透していく。特に心理学における取り込みは早く，Gigerenzer & Murray （1987）によれば，1940 年から 1955 年の 15 年間で「推論革命（inference revolution）」が起こっている。この時期を境にして科学的な方法としての推測統計が必須のものとみなされるようになったのである。Fidler ら（2004）によれば，1955 年にはアメリカの心理学主要雑誌に掲載された論文のおよそ 80% が有意性検定の結果を報告している。日本においても同様で，「心理学研究」の掲載論文における有意性検定の使用率は，1948 年には 4.3%にすぎなかったものが 1955 年には 81.3% にまで増えている（Omi & Komata，2005）。

心理学において最初に導入されたのはフィッシャーの有意性検定であった。しかし，第 2 次大戦後，より完成度の高いネイマン=ピアソン理論が知られるようになると，対立仮説と有意水準を事前設定するという手順や有意水準を第 1 種の過誤の発生頻度であるとするネイマン=ピアソン流の解釈が，応用統計学のテキストに採用されるようになる。また，第 2 種の過誤や検出力などのキーワードも紹介されるようになった。しかしその一方で，全体的な手続きとしてはフィッシャーの有意性検定が科学的推論ツールとして採用され続け，第 2 種の過誤確率も制御する，そのために必要な標本の大きさを事前決定するというネイマン=ピアソンのキーコンセプトは無視された。その結果，図 3 に示すような正体不

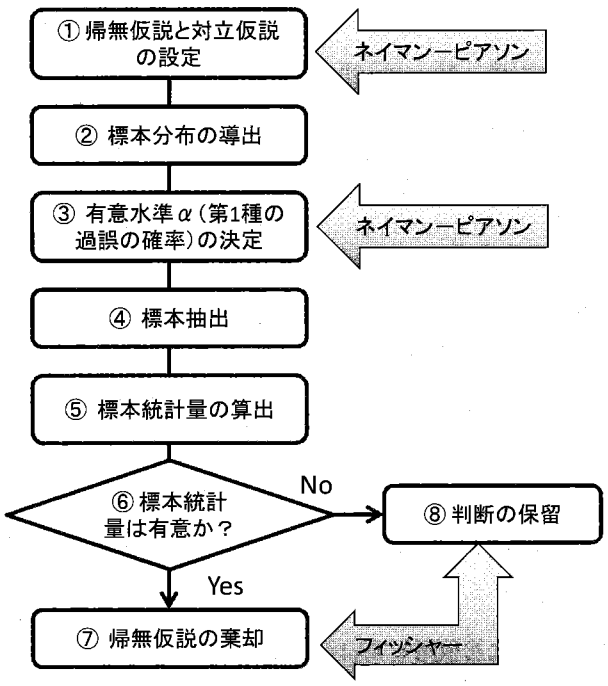

図 3　ハイブリッド仮説検定法

表2　ハイブリッド仮説検定論

	ハイブリッド仮説検定	どちらの立場か
統計的仮説の数	2つ （帰無仮説と対立仮説）	ネイマン=ピアソン*
標本の大きさ	事前決定不要	フィッシャー
α	有意水準＝第1種の過誤確率	ネイマン=ピアソン
β	なし	フィッシャー
$1-\beta$	なし	フィッシャー
意味	相対頻度	ネイマン=ピアソン
αの決定	標本抽出以前	ネイマン=ピアソン
標本抽出	1回	フィッシャー
判定回数	1回	フィッシャー
判定内容	帰無仮説の棄却・判断留保	フィッシャー

*検定仮説として帰無仮説が設定されることが多い。

明のハイブリッド仮説検定法が誕生し，これが唯一無二の「ザ・仮説検定法」として統計ユーザー向けのテキストに記載されるようになったのである。

以下では，Gigerenzer&Murray（1987）にならって，この折衷案を「ハイブリッド仮説検定」と呼ぶことにする。1節で紹介した有意性検定批判はハイブリッド仮説検定に対するものである。

このハイブリッド仮説検定では，検定仮説として効果や差が「ゼロ」であるという帰無仮説を設定することが多く，フィッシャー流に検定仮説を帰無仮説と呼ぶ。ネイマン=ピアソン流に対立仮説や有意水準を事前設定し（図3の①，③），有意水準αは第1種の過誤が発生する頻度であると解釈する。しかし，ネイマン=ピアソンの二分法的な「行動の規則」を最適化するための手続きは含まれていないため（図2の⑤），標本の大きさを事前に決定する必要はなく（図2の②），対立仮説，第2種の過誤，検出力というネイマン=ピアソンのキーコンセプトは統計知識として紹介されるにとどまっている。そのため，ネイマン=ピアソン流に対立仮説と有意水準を事前設定することを要請しているにもかかわらず，検定結果の判断はフィッシャーの方法を採用している。有意でない場合，帰無仮説を採択するの

ではなく判断を保留するのである（図3の⑧）。この手続きに従えば，有意水準を事前に設定さえすれば，データを用いて帰無仮説の「棄却」または「判断の保留」という判定を機械的に行うことができる。

2節で述べたように，フィッシャーは「科学的な帰納的推論」，ネイマン=ピアソンは「最適な決定(帰納的な行動)」が目的であった。では，このハイブリッド仮説検定の目的はなんであろうか？社会科学領域で用いる場合には，科学的な帰納的推論であろう。判断ミスが一定の頻度で発生することを黙認して単一の結果から何かを性急に決定する必要はない。むしろ同様の研究を何度も繰り返し行い，それらの結果を総合して判断することが求められる。しかし，ハイブリッド仮説検定の文脈ではその点があまり強調されない。1回性を前提としたフィッシャーの手続きが土台となっているにもかかわらず，ネイマン=ピアソン的に単一の検定結果から何かを決定できる，つまり，第1種の過誤確率(有意水準)を一定水準以下に抑えて帰無仮説の棄却がなされれば，それを根拠として何かを結論してよいという誤ったニュアンスが漂う。フィッシャーもネイマン=ピアソンも望まなかったに違いない検定方法が誕生してしまったのである（表2）。

実際のところ，互いに相容れない理論を無理やり合体させたこのハイブリッド仮説検定は，結果解釈に間違いが生じやすい。相容れない2つのオリジナル理論における解釈や手続きが，ハイブリッド仮説検定に影のように付きまとい，利用者を混乱させたり誤解させたりするからである。

4. ハイブリッド仮説検定の弱点と弊害

さまざまな論者が「有意性検定」批判，つまりハイブリッド仮説検定批判を行い，膨大な論文が提出されているが，それらの論点は本質的なところでは類似している。ここでは，フィッシャーの有意性検定論とネイマン=ピアソンの統計的仮説決定理論の折衷案として不可避的に生じたと思われる3つの問題点を紹介したい[5]。

(1)　有意確率 p 値の解釈にまつわる様々な誤解

(2)　第2種の過誤確率の無視 ― 検出力の問題 ―

(3)　単一研究による二者択一的判断

[5] 標本抽出や母集団の問題など，有意性検定批判の論点は他にもあるが本稿では割愛する。

4.1 有意確率 p 値の解釈にまつわる様々な誤解

まず，有意確率（p 値）とは何かを再確認しておこう。p 値は，帰無仮説が正しいときに手元の標本と同じかそれ以上の差が生じる確率のことである。したがって，たとえば，p 値が 3%であるということは，帰無仮説が正しい（つまり，差がない）状況で実験を何度も繰り返したとき，手元にあるデータと同じかそれ以上に帰無仮説に反する結果となる頻度は 3%，ということである。

有意確率 p 値にはそれ以上の意味はないのであるが，実際のところ次のような誤った解釈がなされている（Fidler, 2005）。

(1) p 値は帰無仮説が正しい確率である。

(2) 1− p 値は，対立仮説が正しい確率である。

(3) p 値は，得られた結果が偶然の産物である確率である。

(4) p 値は，実験結果の再現性の指標である。p 値が小さいほど，再現性が高い。

(5) p 値は，効果の大きさの指標である。p 値が小さいほど，効果は大きい。

(6) 「統計的に有意でない」ことは，「効果がない」ことを意味する。

(7) 統計的に有意な結果はすべて，理論的にも重要である。

Oakes（1986）の調査によれば，研究者や統計的分析を教える教員でさえ，このような誤解をしている者が少なくない。この状況は現在まで続いており，国内の統計初学者向けに書かれたテキストや解説文ではこれらの誤った解釈に基づく記述が頻繁にみられる。「正確さ」と「わかりやすさ」を天秤にかけ「わかりやすさ」を選択した結果であるのかもしれないが，その弊害は思いのほかに大きい。これについては 4.2 節および 4.3 節で述べる。以下では，誤解（1）～（7）について少し詳しく紹介しよう。

誤解（1）は，A：「帰無仮説が正しいときにある結果が得られる確率」と，B：「ある結果が得られたときに帰無仮説が正しい確率」という 2 つの異なる確率を同一視することに起因する誤りである。仮説検定の枠組みで扱える確率は A のみであり，確率 B を扱うことはできない。確率 B はベイズ統計の逆確率（事後確率）そのもので，その推定には事前確率と条件確率を仮定する必要がある。誤解（2）も（1）と同種の誤りである。仮説検定では，帰無仮説の場合と同様に対立仮説が正しい確率を直接議論することはできない。

誤解（3）は，p 値が有意といえるほど小さくないときに（たとえば，$p = 0.30$），今回得られた結果はたまたま偶然に得られたもので，その偶然性を考慮に入れれば「帰無仮説が正しい」と判断できる，とネイマン＝ピアソン流に解釈してしまうことである。しかし，

このような解釈は，対立仮説が正しいのに帰無仮説を採択してしまうという第 2 種の過誤を無視することにつながる。実際には，母集団における効果や平均値差が，標本誤差と比べてそれほど大きくないときには第 2 種の過誤はかなりの高確率で発生する。

以上の誤解（1）～（3）は，結局のところ，仮説検定では本来的に推定不可能な「あるデータが得られたという条件のもとでの帰無（または対立）仮説が正しい確率」を，p 値のみを用いて評価しようとする誤りである。

誤解（4）は，別の表現を使えば，有意水準 5%より有意水準 1 %で有意といえる結果の方が再現性が高い，偶然の産物である確率が小さく信頼できると解釈することで，有意水準を再現性や信頼性の客観的な指標と解釈するものである。過去にはこの解釈が学術論文の審査に影響を与えたこともある（Bakan, 1967, 訳書 p.230)。しかし，そもそもハイブリッド仮説検定における有意水準は第 1 種の過誤を制御するために事前に設定するものであって，それ以上でもそれ以下でもない。さらに，社会科学領域ではフィッシャーの帰納的推論方法としての有意性検定の概念が浸透しているため，p 値を算出してからその大きさに合わせて有意水準を事後的に選択し，

- p 値が 0.05 未満のとき，「5%水準で有意差あり（$p < .05$)」
- p 値が 0.01 未満のとき，「1%水準で有意差あり（$p < .01$)」
- p 値が 0.001 未満のとき，「0.1%で有意差あり（$p < .001$)」

と報告することで，その標本が帰無仮説からどの程度乖離しているかを示すことが論文報告上の慣行となってきたが，ある特定の標本から算出された p 値はその標本の特性を表す指標でしかない[6]。結果の再現性や信頼性は，標本の大きさや実験デザインに大きく依存しており，p 値だけから測れるものではない。

誤解（5）～（7）は，すべて p 値の大きさと効果の大きさとを直接的に関係付けることによって生じる誤解である。p 値は，標本の大きさと効果の大きさという 2 つの変数から決定されるものであるから，p 値だけから効果の大きさを評価することはできない。有意な差があるということは，その差が実質科学的に意味をもつほど大きい差であることを保証しているわけではなく，その差が取るに足りないほどわずかなものであっても，標本数

[6] 2009 年に出版された APA Manual 第 6 版では，有意水準を事前に設定した上で，各検定結果の報告では p 値そのものを小数点以下第 2 位または第 3 位まで報告することを求めている（ただし，p 値が 0.001 未満の場合は $p < .001$ と表記する)。p 値の大きさに応じて有意水準を決める報告スタイルは，p 値を直接求めることができず，いくつかの有意水準と臨界値とを対応させた数表しか利用できなかった時代の名残であると明言している。

が十分大きければ，$p < .001$ といった高度に有意な水準で帰無仮説を棄却できる。逆に，何らかの理由で標本数を大きくできない場合は，相当の効果がみられたとしても有意であると判断されない場合もある。「統計的有意性」と「効果の大きさ」は別物であり，「統計的有意性」と「実質的な重要性」とを同一視することはできない。

実際のところ，研究者が知りたいのは帰無仮説が正しいのかどうか，効果（差）があると考えられる場合にはその効果（差）はどのようなものなのか，ということであるが，p 値はそれらについての情報を十分には提供しないのである。誤解（1）～（7）は，ハイブリッド仮説検定の枠組みでは本来的に入手不可能な情報を，p 値に無理やり投影させようとすることから生じているといえよう。

なお，ハイブリッド仮説検定では第 2 種の過誤確率を考慮しないのであるから，ネイマン=ピアソン流に「帰無仮説（効果がない）を採択」することは明らかな誤りである（誤解（6））。実際，第 2 種の過誤確率はそれほど小さくない。次節では，この第 2 種の過誤について述べる。

4.2 第 2 種の過誤確率の無視 — 検出力の問題 —

ここであらためて「第 1 種の過誤」と「第 2 種の過誤」を定義しておこう。これらは，ネイマン=ピアソンの二者択一の決定理論において導入された概念である。

第 1 種の過誤は，帰無仮説が真であるのに棄却してしまう誤りのことである。帰無仮説が正しく母集団の平均が同じであっても，それぞれの母集団から無作為抽出した標本の平均値が等しくなることはまずない。標本数が少ない場合には，相当なばらつきが見られるであろう。有意水準 α の検定は，何度も同じ実験を繰り返したとき，帰無仮説が真であるのに偶然の結果の偏りによって誤って帰無仮説を棄却してしまう頻度が α であることを意味している。

第 2 種の過誤は，帰無仮説が偽なのに帰無仮説を棄却しない誤りのことで，いいかえれば，対立仮説が真なのに対立仮説を棄却してしまう誤りである。このような誤りは，母集団間の差がそれほど大きくない場合には，高頻度で発生しうる。たとえ差があったとしても，その差が標本のばらつきの中に埋もれてしまい，検出に失敗するのである。この第 2 種の過誤の発生頻度を β とすると，真実（帰無仮説と対立仮説のどちらが正しいか）と仮説検定の判定結果の成否およびその発生頻度との関係は，表 3 のようになる。2 つの過誤確率 α と β は拮抗する関係にあり，第 1 種の過誤確率 α を小さく設定するほど帰無仮説を棄

表3　仮説検定における真実と判定結果の関係

真実＼判定	帰無仮説を採択（発生頻度）	対立仮説を採択（発生頻度）
帰無仮説が真	正しい判断（$1-\alpha$）	第1種の過誤（α）
対立仮説が真	第2種の過誤（β）	正しい判断（$1-\beta$）

却するための条件が厳しくなるから，第2種の過誤の発生頻度βは大きくなる。

しかし，ハイブリッド仮説検定では，第2種の過誤の発生頻度βを明示的に制御しないため，「母集団に差がある」と帰無仮説を棄却する際の判断の誤りの確率αは有意水準として把握しているが，「有意差なし」と帰無仮説を棄却しない場合にどのくらいの頻度で第2種の過誤が発生しているかが明らかでない。そのため，この検定法では「帰無仮説を採択」することはできず，「判断の保留」を行う（図3)。ところが，多くの有意性検定批判で指摘されているとおり，また，4.1節で述べたように，「帰無仮説を棄却しない＝帰無仮説を採択」と解釈する機械的な二者択一的解釈が浸透しており，実際のところ，多くの入門的なテキストや解説文で「帰無仮説を採択する」という表現が採用されている。

では，第2種の過誤はどのくらいの頻度で発生するのであろうか。検定力分析という手法を用いると，この頻度を比較的容易に推定できる。検定力（以下では検出力と呼ぶ）とは，「対立仮説が真であるときに帰無仮説を棄却する」という正しい判断を行う確率，つまり，差を検出する確率のことで，$1-\beta$である。検定力分析を用いると，有意水準α，検出力$1-\beta$，効果量および標本数の関係を調べることができる。

ここで効果量とは，標本数によらない実質的効果の大きさの表す重要な指標である。平均値差に対してはグループ毎の平均値差を標準化したもの，変数間の関係の強さに対しては標本相関係数や連関係数などが用いられる（詳しくは，Cohen, 1988；水本＆竹内, 2008)。グループ毎の平均値差を検討する場合，その差が標準偏差と同じ程度であるときの効果量（ここでは，標準化された平均値差）の値は1であり，標準偏差の半分程度の差であれば0.5となる。

では，第2種の過誤の発生頻度βの話に戻ろう。検定力分析では，βではなく検出力$1-\beta$を用いて議論されるので，ここでもそれに倣って標本数と検出力との関係を表4に示す（Cohen, 1988; 橘 1986)。有意水準5%，両側検定で独立な2群の平均値差について検定

表4　t検定における検出力，効果量，標本数の関係（橘，1986）

	標本数	効果量		
		小効果量 (0.2)	中効果量 (0.5)	大効果量 (0.8)
有意水準 $\alpha=0.05$	8	0.07	0.15	0.31
	20	0.09	0.33	0.69
	50	0.17	0.70	0.98
	100	0.29	0.94	*
	200	0.51	*	*
	1000	0.99	*	*

＊は検出力が0.995以上であることを示す。

する場合（t検定）の例である。表中の標本数は各群ごとの値である。標本数と効果量を決めると，その検定の検出力を求めることができる[7]。たとえば，各群の標本数が50，効果量が0.2と小さい場合には，検出力はわずかに0.17である。母集団において効果量にして0.2という差があっても，有意となる確率は0.17，つまり100回のうち17回しか有意とならず，83回は差の検出を失敗する（第2種の過誤が発生する）のである。また，効果量が0.5と中程度の効果がある場合でさえ検出力は0.70であり，10回に3回は差の検出を失敗することがわかる。

研究分野によって多少の差はあるものの，心理学分野の論文誌に発表されている研究の平均的な検出力の水準はそれほど高くないことが示されている（Sedlmeier&Gigerenzer, 1989）。杉澤（1999）の，1992年から1996年までに発行された『教育心理学研究』掲載論文の検出力調査においても，中程度の効果量の研究に対する平均検出力は0.66，小効果量の場合には0.23であることが報告されており，実際の研究場面においてかなりの頻度で第2種の過誤が発生していることが示唆される。

検出力への配慮を怠るとさまざまな問題が生じうる（Cohen, 1994; 杉澤, 1999; 橘, 1986）。

[7] 「効果があるかどうか」を知るために研究をおこなうのだから，母集団における効果量の大きさを事前に想定するのは難しい。代用として標本効果量を用いるという方法もある（豊田, 2009）が，効果量を大きく見積もりすぎる可能性がある。

たとえば，検出力が低い研究が多くなされると，理論的発展が期待できる新たな発想が，「有意な結果が得られない」，「有意でないことが多い」という理由のために未発表のまま葬られてしまう可能性がある。とくに，予備的研究段階では標本数が少ないことが多いから，より影響が大きいと思われる。逆に，標本数が多く検出力が高すぎると，それほど本質的でない差であっても高度に有意であると検出され，非常に小さな p 値を結果の再現性や信頼性，あるいは効果の大きさの指標であると誤って解釈してしまうと（4.1 の誤解（4），（5）），Bakan（1967）が指摘するように「『有意な』結果の掲載は，それから先の研究を停止させる」ことになる。

実際，検出力の低い研究において有意な結果が得られなかった場合に「有意でない」＝「効果がない（差がない）」と帰無仮説を採択してしまうという解釈の誤りが（誤解（6）），理論の発展に少なからぬ影響を与えることがある。次節では，検出力への配慮不足のために実際に生じた研究の混乱の事例を，Fidler（2005）の 3 章 “Has NHST Dameged Science?” から抜粋して紹介しよう。

4.3　単一研究による二者択一的判断

ネイマン=ピアソンの仮説決定理論はそもそも繰り返し実験を前提としているが，「一回性」を重視する立場のフィッシャーも，科学的推論における繰り返し実験の重要性を説いている。少し長くなるが，フィッシャー（1966）が「実験計画法（“The Design of Experiments”）」第 2 章（実験の原理）の中で述べている言葉を引用しよう。

> “自然に関する知識の進歩”の中では，すなわち，経験ないしは計画された一連の実験から学ぶときには，結論はつねに暫定的なものであって，それまでに得られた証拠を解釈してそれを一体化した経過報告という性質のものである。5%，2%，または 1%という慣行上の有意水準を用いて，仮説が矛盾するという注釈をつけるのが便利であるとしても，帰納的推論において，その証拠が実際に到達している正確な強度をみおとしたり，その後の試行によって，それがさらに強くなり弱くもなりうることを無視したりしてよい，ということには決してならないのである。…… 純粋な研究の分野では，誤った結論による出費とか，一そう正確な結論に到達するのが遅れることによる出費の評価は，考えられるところでは，見せかけのもの以上ではありえないし，またこのような評価はいずれにしても，科学上の証拠の状態を判断する上では，承認

しがたくまた不適切なものである。(Fisher, 1966, 訳書 p. 21－22)

このようにフィッシャーも帰納的推論において単一の実験から何かを結論することには否定的である。以下は，個別研究において二者択一的な判断が行われ，それらの結果が統合されないまま検定結果の不一致（有意かどうか）によって混乱が生じ，学問や臨床への応用が停滞した事例である。統計的仮説検定がこれまでの実証研究の発展を支えたことは事実であり，ここで紹介するような研究の混乱は，一種の副作用のようなものである。また，各研究の効果量をもとに研究結果を統合するメタ分析や検出力という概念の普及によって，この副作用の発生は格段に減少しつつある。しかし，p 値や検出力についての理解が不十分なまま仮説検定を用いて二者択一的判断を行うことの危うさを確認するために，簡潔に紹介しよう。

Fidler が最初に挙げているのは，「状況特異的妥当性理論（the theory of situation-specific validity)」である。新卒学生の採用試験などで広く用いられている職業適性検査は 20 世紀初頭にその開発が始まっているが，最初の数十年間は検出力への配慮不足のために大きく混乱したという。同じ方法を用いてさまざまな職場における適性検査の予測的妥当性を調べると，同種の仕事や同じ会社が対象の場合でさえ，職場ごとにまったく異なる結果が得られたのである。1930 年から 1940 年代には，このような状況が生じる原因は顧客の違いや上司の指導スタイルの違い，研修内容の違いなど，研究者が制御しきれない微妙な要因の差であると結論された。職業適性結果の予測的妥当性は職場の個別的状況によって異なるという理論が出来上がったのである。

しかし，1970 年代の後半に，Schmidt & Hunter(1998)が複数の研究結果を統合するメタ分析を用いて，これらの一見相矛盾する不一致は検出力不足に起因して人工的に作り出された錯覚であることを明らかにした。ほとんどすべての研究において効果量は同じ傾向を示していたが，各研究の標本数は 40～70 と少なかったために十分な検出力が得られず，研究ごとに「有意」であったり「有意でない」という結果が得られたりしていたのである。

この他にも，検出力への配慮不足と二者択一的な仮説選択によって見せかけの結果の不一致が生じ，理論の発展および臨床への応用が遅れた事例として，実験心理学者のセリグマン（Seligman, M.E.P.）が提唱した学習性無力感と抑うつに関する理論や，急性心筋梗塞に対するストレプトキナーゼ静脈内投与の有効性に関する研究など 6 つの事例が紹介されている。現在ではストレプトキナーゼに代表される血栓溶解薬が急性心筋梗塞の治療に有効であることは広く知られているが，この薬の有効性を確認するのに，1959 年から 1988

年までの 30 年間，33 もの臨床試験が費やされている。Fidler は，これらの臨床試験結果が適切に統合されていれば，1988 年より 15 年も早く，1973 年には有効性が確認できた可能性があると述べている。

5. アメリカ心理学会における統計改革 — おわりにかえて

Cohen の有意性検定に対する批判論文“The earth is round ($p < .05$)”が契機となって始まった「心理学における統計改革」は現在どのような状況にあるのだろうか。APA の推測統計に関する専門委員会が 1999 年に提出したガイドラインでは，分析結果の報告に関して次のような指針を策定している（Wilkinson ら，1999）。

a) 母集団における効果量をどのように見積り，どの程度の検出力を想定しているかなど，標本の大きさを決定したプロセスを明記すること。結果報告においては，事後的に標本検出力を算出することは避け，その代わりに信頼区間を用いること。

b) 帰無仮説の棄却または採択という二者択一的な検定結果を報告することが，実際の p 値や信頼区間を報告することより優れているという状況は想像しがたい。また，「帰無仮説を採択する」という不適当な表現を決して使わないこと。主要な結果に対して，また p 値を報告する場合は必ず，（標本）効果量を報告すること。効果量の報告は，将来の検定力分析やメタ分析に有用な情報を与える。

c) 効果量は，信頼区間を用いた区間推定結果を示すこと。

ここで信頼区間（confidence interval, CI）とは，ある信頼水準のもとで標本から推定される，真の値（母数）を含むと考えられる区間のことで，95%信頼区間や 99%信頼区間がよく用いられる。95%信頼区間とは，標本抽出および信頼区間の算出作業を 100 回繰り返したとき，それぞれの信頼区間に母数が含まれる頻度の期待値が 95 回であるという意味である。したがって，ある研究において母平均の 95%信頼区間が 50 から 60 までと推定されたとすると，その推測が的中している確率は 95%ということになる[8]。なお，有意水準 5%の仮説検定で効果が 0 かどうかを検定することは，その効果の 95%信頼区間に 0 が含まれているかどうかと同値であるので，信頼区間を報告することによって仮説検定において帰無仮説が棄却されるかどうかという情報も伝えることが出来る。それに加えて，効果の大きさがどの程度であると推定されるか，推定誤差はどの程度かという，より実質的な情報を

[8] 母平均が 50 から 60 までにある確率が 95%ではない。

提供できるのである。

APA の専門委員会の報告は，区間推定（信頼区間）を用いて効果量を推定することの必要性，統計的有意性と理論的有意性を区別することの重要性を強調した内容になっており，有意性検定一辺倒の姿勢から，信頼区間を用いた推定重視へ，いいかえれば，「効果があるかどうか」の判定から「効果の大きさはどの程度か」の推定へと大きく移行した内容となっている。

このような専門委員会の提案を受けて 2001 年に改訂された APA Publication Manual 第 5 版は，効果量，エラーバーを含む図，信頼区間による報告を推奨し，特に信頼区間の利用を「最適な報告戦略（the best reporting strategy）」であると述べている。この論文執筆マニュアルは，心理学およびその周辺分野の学術雑誌においても手本とされることが多いため，その改訂が社会科学研究に与える影響は大きく，このような APA 主導の一連の変化は「心理学における統計改革（statistical reform）」と呼ばれている。

しかし，第 5 版では効果量や信頼区間の報告は義務付けられず，マニュアルに掲載されている具体的な使用例や記述例は旧来の仮説検定による方法がそのまま残り，マニュアルが推奨する「最適な報告戦略」の具体的な使用例は提示されなかったため，統計改革の推進者たちを大きく落胆させる結果となった。Cumming ら（2007）の調査によれば，1998 年以降の心理学系学術雑誌に掲載された論文において効果量や信頼区間を用いた報告は増加しつつあるものの，ほとんどすべての論文において仮説検定結果が合わせて報告され，結果の解釈は旧来どおりのものが多い。Cumming らはこの現状を“Change, but little reform yet”と表現している。Fidler ら（2004）は，医学誌では 1980 年代に信頼区間への移行（統計改革）が迅速に完了したことと対比させて，心理学における統計改革の推進に必要なものは，テキストの改訂，利用方法に関する詳しいガイドラインと手本となる実践例，論文誌編集者や学界の強いリーダーシップであると指摘している。

2009 年，APA Publication Manual 第 6 版が幅広い行動科学・社会科学領域の学生・研究者を対象として出版された。今回は，大幅改訂が必要な重点領域として統計が指定され，ワーキンググループが設置された。そのメンバーには統計改革の推進者の一人である Cumming が加わり，統計改革を強く推し進める内容に改められた。検出力の重要性が明記され，また，効果量と信頼区間による結果報告を強く推奨するとともに具体的な報告方法が豊富な例によって示されている。APA Manual 第 6 版では統計的仮説検定について次のように言及している。

歴史的に心理学研究者は，多くの（しかし，全てではない）分析的なアプローチの出発点として帰無仮説有意性検定（NHST）を強く信頼してきた。APA は，NHST が単に出発点でしかないこと，分析結果の意味するところを完全に伝えるためには，効果量，信頼区間，詳細な記述を付加して報告することが必要であることを強調する。各論文誌が NHST を重視する程度は，それぞれの編集者が決めることである。しかしながら，検定を行ったすべての仮説と適切な効果量および信頼区間の推定結果を報告することは，すべての APA 論文誌にとって最低限の期待である。科学研究者は，研究結果を正確かつ信頼できる形で報告することに対して常に責任を負っている。(APA，2009，p.33，筆者訳)

さらに 2010 年には，社会科学領域の論文査読者のための量的方法のガイドラインが出版され，効果量と信頼区間に関する詳しい解説が掲載されている（Cumming&Fidler，2010)。ようやく，論文執筆者と査読者の双方に今回の統計改革が広く浸透し実質的な実践が広まる準備が整いつつある。

推測統計学は，20 世紀の前半以降も進歩をつづけており，例えばネイマンは統計的仮説決定理論よりもむしろ区間推定を重視していたという（木村，1992)。しかし社会科学においては，旧来の欠点の多い方法が化石のように固化して依然として使われ続けてきた。日本の心理学分野における統計改革の状況について大久保 (2009) が詳しく調査しているが，効果量や信頼区間が報告されることはほとんどなく，進んでいないことが報告されている。

Rozeboom（1962）は，本稿の冒頭で引用した文章に続いて次のように述べている。

伝統的な帰無仮説の手法が，現在の統計理論のなかでより満足ゆく一連の推論技術に道を譲ってからすでに久しい。しかし，とくに方法論の問題を論ずるときには，心理学者の感覚的な防衛が効果を発揮しており，このため，比較的初期のころの統計的思考様式が局所的に支配をつづけている。(Morrison&Henkel，1970，訳書 p.204)

半世紀前の 1960 年代とは異なり，統計ソフトウェアを使えば検出力は容易に計算できる。標準的な統計処理ソフトウェアは信頼区間の算出機能を有している。また，APA が推奨する効果量や信頼区間は本質的に仮説検定と同じ考えに沿ったものであるから，今回の統計改革の普及に大きな障壁やデメリットがあるとは考えにくい。さらに，信頼区間は直感的にわかりやすく，本稿で指摘したような誤った結果の解釈もおこりにくい。さらに，効果量は検定力分析やメタ分析に不可欠のものであり，これを報告しないことはその研究の価

値を大きく損なう。

先に述べたように，アメリカ心理学会ではようやく統計改革が浸透する準備が整いつつある。Rozeboom が指摘したような心理的防衛によって，この統計改革の流れが日本において停滞するとすれば，大変残念なことである。APA Manual 第 6 版の翻訳が待たれるのは当然のことながら，それに呼応した教育用テキストの改訂，研修や教育プログラムの充実が強く望まれる。

なお，統計改革は今回が終わりとは限らない。有意性検定に代わる手法として提案されるものの 1 つにベイズ統計がある（たとえば，Rozeboom，1960; Gigerenzer&Murray，1987; Kline，2004; 松原，2008; Läärä，1999）。ベイズ統計では，科学者の真の関心の対象である「観察された実験結果のもとで自分の仮説が正しい確率はどの程度か」という確率を一貫した論理で直接扱うことができため，科学的な推論との親和性が非常に高い。「結果（データ）から原因を考える」というごく自然な人間の思考方法を体系化した統計理論であるが，人間が考える主観的な信念あるいは確信の度合など客観的には求められない量をも確率と解釈しその理論に取り込むことから，客観性を重視するフィッシャー学派やネイマン=ピアソン学派とは相いれず，1920 年代に強く批判され衰退した。しかし，1950 年代にベイズ理論として再構築され，今日の数理統計学において確固とした地位を築いている。近年，国内においてもベイズ統計学の入門テキストの出版が増え（たとえば，松原, 2008），また，生態学研究ではすでにその取り込みが始まっている（McCarthy, 2009）。社会科学における統計改革の第 2 段階はベイズ統計学への移行かもしれない。

しかし，研究遂行において最も重要なものは，研究理論とそれに基づく研究者の論理的思考であり，どのような道具を用いるかではない。本稿では統計的仮説検定を不十分で使いにくい分析道具と位置付けたが，研究者がその道具を十分に理解し使いこなすことさえできれば，実はどの道具を使うかはそれほど重要なことではない。Wilkinson ら（1999）が APA のガイドラインで指摘したように，データ分析にも「オッカムの剃刀」の原理が当てはまり目的に応じた必要最小限の分析を行うことが望ましいのである。どのように優れた道具が出現したとしても，研究者がその原理を理解しないままブラックボックスから出力された結果を鵜呑みにするような状況であれば，そのような道具を使うべきではない。統計改革におけるもう 1 つの重要なテーマは，研究道具の 1 つに過ぎない統計手法とそれを使いこなすためのハウツーへの過度の依存から研究者が脱却することにあるのではないだろうか。

引用・参考文献

American Psychological Association (2009) *Publication Manual of the American Psychological Association* (6th ed.), Washigton, DC, Author.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Psychology Press.

Cohen,J.(1994) The earth is round ($p < .05$) , American Psychologist, 49, 997-1003.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J. McMenamin, N. and Wilson, S. (2007) Statistical Reform in Psychology: Is Anything Changing?, Psychological Science, 18, 230-232.

Fidler, F. (2005) From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology (Doctoral dissertation, Department of History and Philosophy of Science, University of Melbourne), Retrieved from https://www.latrobe.edu.au/psy/staff/fidler.html.

Fidler, F., Cumming, G, Burgman, M. and Thomason, N. (2004) Statistical reform in medicine, psychology and ecology, Journal of Socio-Economics, 33, 615-630.

Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*, Oliver and Boyd（渋谷政昭・竹内啓訳『統計的方法と科学的推論』，岩波書店, 1962）.

Fisher, R.A. (1966) *The Design of Experiments* (8th ed.), Oliver and Boyd（遠藤健児・鍋谷清治訳『実験計画法』，森北出版, 1971）.

Gigerenzer, G. and Murray, D.J. (1987) *Cognition as Intuitive Statistics*, Lawrence Erlbaum Associates, Inc.

Hancock, G.R. and Mueller, R.O. (eds.) (2010) *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, Routledge.

細谷雄三（2002）『統計的証拠とその解釈　増補版』，牧野書店.

木村和範（1992）『統計的推論とその応用』，梓出版社.

Kline, R.B. (2005) *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research,* American Psychological Association.

Läärä, E. (2009) Statistics: reasoning on uncertainty, and the insignificance of testing null, Annales Zoologici Fennici, 46, 138–157.

Loftus, G.R. (1996) Psychology will be a much better science when we change the way we analyze data, Current Direction in Psychological Science, 5, 161-171.

松原望（2008）『入門ベイズ統計学―意思決定の理論と発展』，東京図書.

McCarthy, M.A.（2009） 野間口 眞太郎（訳）『生態学のためのベイズ統計』，共立出版

Morrison, D.E. and Henkel, R.E. (eds.) (1970) *The Significance Test Controversy*, Chicago, Aldine.（内海庫一郎・杉森滉一・木村和範訳『統計的検定は有効か―有意性検定論争―』，梓出版, 1980.）

芝村良（2004）『R.A.フィッシャーの統計理論』，九州大学出版会.

水本篤・竹内理(2008)研究論文における効果量の報告のために―基礎的概念と注意点―，英語教育研究, 31, 57-66.

Oakes, M. (1986) *Statisical Inference: A Commentary for the Social and Behavioural Sciences*, John Wiley & Sons.

大久保街亜（2009）日本における統計改革―基礎心理学研究を資料として―，基礎心理学研究 28, 88-93.

Omi, Y. and Komata, S. (2005) The evolution of data analyses in Japanese psychology, Japanese Psychological Research, 47, 137-143.

Schmidt, F.L. and Hunter, J.E. (1998) The validity and utility of selection methods in personnel psychology: practical and theoretical implication of 85 years of research findings, Psychological Bulletin, 124, 262-274.

Sedlmeier, P. and Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies?, Psychological Bulletin, 105, 309-316.

杉澤武俊（1999）教育心理学研究における統計的検定の検定力，教育心理学研究，47, 150-159.

橘敏明（1986）『医学・教育学・心理学にみられる統計的検定の誤用と弊害』，医療図書出版.

豊田秀樹(編著)（2009）『検定力分析入門―R で学ぶ最新データ解析―』，東京図書.

Wilkinson, L. and the Task Force on Statistical Inference (1999) Statistical methods in psychology journals: Guideline and explanations, American Psychologist, 54, 594-604.