The Annual Report of Educational Psychology in Japan
1995, Vol. 34, 122-131

展　望

教育心理学研究と統計的検定

南風原　朝　和
（東京大学）

STATISTICAL TESTING IN EDUCATIONAL AND PSYCHOLOGICAL RESEARCH

Tomokazu HAEBARA

This article is intended to provide an overview of the methodological issues in the use of statistical tests in educational and psychological research. First, an exposition is made of those basic aspects of the theories which are not covered in typical applied statistics textbooks and which are often misunderstood. Then such recent efforts to improve the use of statistical tests as power analysis and effect-size estimation are summarized. The final section of the article discusses issues concerning the assumptions and interpretations of statistical tests, generalization of research results, and statistics education for prospective researchers.

Key words : statistical significance, statistical power, effect size, generalization, statistics education.

本論文は，教育および心理学研究における統計的検定の利用に関する方法論的問題を概観することを目的としている。まず，統計的検定の理論のうち，典型的な心理統計学の解説書では触れられていない基本的な側面，しばしば誤解される側面について解説を行う。そして，検定力分析や効果の大きさの推定など，統計的検定の利用を改善するために提案されてきた方策についてまとめる。最後に，統計的検定の仮定と解釈，研究結果の一般化，そして将来の研究者のための統計教育の問題について議論を展開する。

1. はじめに

最近の日本の心理学における統計的方法の利用の現状を調査した尾見・川野（1994b）によると，1992年3月から1993年3月までに発行された『教育心理学研究』に掲載された論文62編で使用されている統計的方法の上位5種類は，以下の通りである。

1位	分散分析	40編
2位	ティ検定	31編
3位	積率相関係数	27編
4位	カイ2乗検定	17編
同位	因子分析	17編

同じ時期に発行された『心理学研究』でもこの順位は全く同じであり，1992年4月から1993年4月までに掲載された論文68編のうち，これらの方法を使用しているものは，上から順に，45編，33編，23編，13編，11編となっている。

このうち，1位の分散分析と2位のティ検定は，いずれも平均値の比較に統計的検定を利用するものである。3位にランクされている積率相関係数も，そのほとんどが統計的検定を伴っていると思われる。また，4位のカイ2乗検定は，カテゴリーデータの連関の評価に統計的検定を利用するものである。このように，心理学の研究においては統計的検定が非常に多く用いられており，上記の期間に発表された『教育心理学研究』の論文のうち統計的検定を用いていないのは，わずかに6編だけであり，『心理学研究』では3編にすぎない。（ここまでの文章を読んで，「両雑誌の間の“検定論文”の比率の有意差は？」と尋ねたくなった読者も多いのではないだろうか。後述の4の(1)参照。）

この数は，学会誌に投稿され採択された研究論文に関するものであり，統計的検定がそのような研究において特に多用されているという可能性もある。しかし，それにしても，統計的検定なくして心理学研究は成立しないかのような圧倒的な数である。「研究計画は分析方法まで考慮して決めること」というのが，研究指導における決

まり文句だが，このような状況では，分析方法としてまず統計的検定が選ばれ，それを前提とした研究だけが計画されるということになりかねない。

一方では，統計的検定自体の限界や，統計的検定を利用する際の誤解・誤用の指摘が，これまで繰り返しなされてきた。1940年代から1960年代までに社会学と心理学を中心に展開された検定批判論の主なものは，Morrison & Henkel (1970) に収録されており，その抜粋 (30編中の17編) は内海らによって1980年に『統計的検定は有効か』というタイトルで翻訳・出版されている。また，単独の著者による首尾一貫した批判論としてはOakes (1986) がある。わが国でも橘 (1986) の『医学・教育学・心理学にみられる統計的検定の誤用と弊害』が出版されており，統計的検定にかかわる問題点は，これらの著書でほぼ尽くされていると言ってよいだろう。

しかし，学会での研究発表や各種の研究報告を見る限り，統計的検定の基本原理や問題点に関する理解が広く浸透しているとは思われない。特に，統計的検定に(誤解に基づく) 過大な期待を抱き，その結果として誤った結論を導き出したり，検定結果を左右する要因についての知識が足りないために効率の悪い研究を繰り返したり，といったケースは少なくない。

こうした現状を改善する上で，誤用の実例の指摘(長谷川，1994b) や具体的なガイドラインの提供 (繁桝，1994) は有用であろう。しかし，誤用であるか否かの判断は必ずしも容易でない場合があり，また，具体的なガイドラインについても，専門家の間で意見が食い違うこともある。専門家の助言を参考にするにせよ，結局は，個々の研究者自身の知識に基づいて，それぞれの研究の文脈に応じた判断をするしかないのである。

本稿は，研究者のそうした主体的で知的な判断の一助となることを目的としてまとめた，統計的検定に関する「解説的展望」である。まず，標準的な心理統計学の教科書では十分な解説がなされていない基本事項を中心に，統計的検定の理論について述べる。次に，検定利用の改善のために提案されてきた方法を，その改善案自体に内在する問題とともに紹介する。そして最後に，いくつかの方法論的問題と統計教育の課題について議論を展開したい。

2. 統計的検定の理論について

(1) Fisher 理論と Neyman-Pearson 理論

統計的検定の理論は，R.A.Fisher によって「帰無仮説検定」の理論，あるいは「有意性検定」の理論として確立され，Fisher (1925, 1935, 1956) などの著作を通して，心理学を含む広範な領域の研究者に利用されるに至った。しかし，対立仮説，2種類の誤り，検定力など，現在の心理統計学の教科書に登場する主要な概念は，Fisher が激しい論争を繰り広げた相手である J.Neyman と E.S. Pearson の「統計的仮説検定」の理論 (Neyman & Pearson (1933) など) において提唱されたものであり，Fisher 自身はそれらの概念を科学的研究にとって意味のないものとして批判している。

これら2つの立場の違いの一端を，「有意水準はデータ収集に先立って決めるべきか？」という問題を例にとって示すと次のようになる。Fisher にとっては，p 値(次節参照)は，帰無仮説にどの程度の信頼がおけるかを示すものであり，あらかじめ .05や .01という基準を設定し，それによって有意か否かの判断を機械的に行うというようなものではなかった。これに対し Neyman らは，帰無仮説が真であるか否かというような推論ではなく，2つの対立する仮説の間の実際的な選択 (たとえば，工場の生産を一時停止するか否かといった行動決定を伴うもの) の問題として仮説検定を考えた。そして，第1種の誤りの確率 α(有意水準)と第2種の誤りの確率 β を，誤った決定に伴う損失の大きさを考慮してあらかじめ設定し，それに必要な大きさの標本をとって検定を行うことを提唱した。したがって，上記の問題に対しては，Neyman & Pearson の立場に立てばイエス，Fisher の立場に立てばノーという答えになる。

また，Neyman らは，明確に規定された母集団から標本が繰り返し抽出される過程で，誤った判断に伴う損失を最小にするという決定理論的な枠組みを採用したが，Fisher はそれを科学的研究にふさわしくない考え方として批判している。Neyman らは工場での品質管理のように，実際に標本抽出の繰り返しや損失の評価のできる状況を想定して理論を展開し，Fisher は1つの標本のデータに基づいて，仮想的な母集団における帰無仮説の信憑性を評価することを考えていたことからすれば，この対立は当然と言える。(Fisher と Neyman & Pearson の間の，そして E.S.Pearson の父である K.Pearson をも巻き込んだ対立については，安藤 (1989) や Cowles (1989) の叙述がおもしろい。また，Fisher (1925, 1935, 1956) のそれぞれ第14版，第8版，第3版を1冊にまとめたものが，Fisher (1990) として出版されている。)

Gigerenzer (1993；Gigerenzer & Murray, 1987) は，ほとんどの心理統計学の教科書においてこれら2つの相対立する理論が折衷され，あたかも検定の理論が万人に認められた単一の理論であるかのような印象を与えてきたことを指摘し，このことが，検定の無批判的な受容の原因の1つとなったと述べている。そして，検定の理論を作り上げた人々の間に激しい論争のあったこと，さらには統

計的検定とはその基礎からして大きく異なるような推論の形式（ベイズ推論など）もあることを正しく伝えることが，機械的・儀式的な検定利用から研究者を解放するのに必要であるとしている。

(2) 確率の解釈

前節で述べたFisher流のp値の解釈は，あたかも帰無仮説が真である確率がp値によって与えられているかのようだが，仮説の真偽に対して確率を考えるということは，Fisherも Neyman & Pearsonも認めていない。「データに基づいて仮説の確からしさを吟味する」という言語表現を条件付き確率に置き換えれば，P（仮説｜データ）を評価するということになるが，このように仮説自体の確率を考えるのはベイズ統計学の立場である（繁桝，1985）。非ベイズ的な統計的検定では，その逆の形の条件付き確率P（データ｜仮説）しか問題にしない（できない）。

たとえば，検定統計量の値が大きいときに帰無仮説を棄却する片側検定を用いると，Neyman & Pearsonの第1種の誤りの確率αは，確率変数である検定統計量をT，帰無仮説H_0の棄却の限界値をt_cとしたとき，

$$\alpha = P(T \geq t_c \mid H_0)$$

と表わされる。（本稿では，簡単のため，特に断らない限り片側検定を想定して論を進める。）また，p値は，実際のデータで得られた検定統計量の実現値をt_0とすると，

$$p = P(T \geq t_0 \mid H_0)$$

で与えられる。

統計的検定で用いられる確率αやpが，このように帰無仮説を条件としたときのデータの得られ方に関する条件付き確率であるにもかかわらず，ベイズ的に仮説の確率と解釈してしまう誤りは，これまで繰り返し指摘されてきた（最近では，Pollard（1993）など）。しかし，条件的推論そのものの難しさ（Evans, 1989）に加えて，確率的な要素が入り込むのであるから，こうした誤りの除去は容易でない。たとえば，「帰無仮説を棄却したことが誤りである確率はαである」という解釈はしばしば見られるが，「帰無仮説を棄却したことが誤りである」というのは「帰無仮説は正しい」ということと同じであり，結局「帰無仮説が正しい確率はαである」と言っていることになる。したがって，仮説を条件としたときのデータに関する条件付き確率ではなく，仮説自体の真偽にかかわる確率的表現となってしまっている。

データのほうを条件とした解釈の極端なケースとして，等平均，無相関，あるいは分布の正規性といった帰無仮説が棄却されなかったときに，平均の等しさ，相関の無さ，あるいは分布の正規性が「認められた」と解釈することが挙げられる。この場合は，データをもとに，仮説の真偽に関して確率1（または0）の絶対的な評価を下していることになる。なお，帰無仮説の棄却ではなく，採択に関心がある場合の統計的検定については，3の(2)で触れる。

統計的検定で扱われる確率については，その条件性の他にも特筆すべきことがある。それは，先の式から明らかなように，データの確率とはいえ，得られたデータそのものの出現確率ではなく，そのデータよりも極端な（したがって実際には得られていない）データまで含めての確率になっていることである。このことは，ベイズ的にデータを条件として仮説の確率を論じる際に，得られたデータそのものが条件付けの事象となることと対照的である。なお，統計的検定におけるp値と，帰無仮説のベイズ的事後確率の間の関係については，Casella & Berger（1987）と Berger & Sellke（1987）の間の議論がある。

(3) 検定統計量の成り立ちと検定力

検定によって有意な結果が得られる確率，すなわち検定力が，標本の大きさ（被験者数）によって影響を受けるということは，広く認識されているようである。しかし，標本の大きさの決定がしばしば明確な根拠もなくなされている現状（3の(1)参照）からすると，標本の大きさと検定力との間にある直接的な関係が十分には理解されていないかもしれない。

Rosenthal & Rosnow（1984）が研究法に関する著書の中で一貫して示しているように，検定統計量は

検定統計量＝効果の大きさ×標本の大きさ

という形で構成されている。つまり，効果の大きさが同じならば，あとは標本の大きさだけで検定統計量の値，したがって検定の結果が決まるのである。例として，2群の平均値差の検定のための統計量t，2変数間の相関係数の検定のための統計量t，そして，2つのカテゴリー変数間の連関の検定のための統計量χ^2を，上記の式の形で表わしておこう。

$$\text{平均値差の検定：} t = \frac{m_1 - m_2}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$\text{相関係数の検定：} t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

$$\text{連関の検定：} \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(p_{ij} - p_i p_j)^2}{p_i p_j} \times n$$

効果の大きさは，それぞれの検定に特有な形で表わされる。平均値差の検定の場合は，平均値差を各群共通であると仮定される母集団標準偏差の推定値で割った「標準化された平均値差」（本稿では芝・南風原（1990）にならって，これを「効果量」と呼んでおく）が効果の大きさの指標となっている。相関係数の検定の場合は，一方の変数において1標準偏差の差がある2つの群を想定し，他方の変

数についてその2群間の効果量を考えれば，上式に示したような形になる。また，連関の検定の場合は，カイ2乗統計量を算出する式において，度数の代わりにそれぞれを全体の標本の大きさ n で割った相対度数を代入したものが，効果の大きさの指標となっている。

なお，このように定義される効果の大きさを一般に ES とし，標本で得られた ES の値を ES_0 とすると，先に示した p 値の定義式は

$$p=P(ES\geqq ES_0 \mid H_0,\ n)$$

のように書くこともできる。標本の大きさ n を決めれば，効果の大きさが直接的に p 値に反映されるということである。p 値が小さいときに，それをすぐに効果の大きさに結びつけて考えてしまう傾向がしばしば指摘されるが，これは条件付き確率の言葉でいえば，「標本の大きさという条件付け事象を無視してしまう傾向」ということになる。

3. 検定利用の改善の試み

(1) 検定力分析

有意水準，すなわち第1種の誤りを犯す確率 α を固定すれば，検定力，すなわち帰無仮説が棄却される確率 $1-\beta$ は，

$$1-\beta=P(T\geqq t_c \mid \delta)$$

で与えられる。ただし，T は検定統計量，t_c は棄却の限界値，δ は母集団における効果の大きさを表わす。

先に述べたように，Neyman-Pearson 流の仮説検定においては，有意水準だけではなく，検定力もあらかじめ定め，それに必要な大きさの標本を用いて検定することになっている。しかし，1962年の Cohen の報告によれば，その頃までの心理学の研究では，このような方法で標本の大きさを決定しているケースはほとんどなかった。彼は1960年発行分の *Journal of Abnormal and Social Psychology* 誌に掲載された論文のうち，統計的検定を適用している70編について，それらの研究で用いられた大きさの標本でどの程度の検定力が確保されているかを調べた。母集団における効果の大きさについて，大(効果量にして1.00)，中 (同0.50)，小 (同0.25) の3段階を設定したとき，有意水準 .05の両側検定の検定力は“大”効果の場合に平均 .83，“中”効果の場合に平均 .48，“小”効果の場合に平均 .18という結果になった。つまり，平均的に言って，“小”効果を検出できる可能性は非常に低く，“中”効果の場合も50%に満たないということである。多くの研究における統計的検定がこのような低い検定力で行われていると，理論的な発展が期待できるような重要な発想も，統計的有意性が得られないために人目に触れることなく葬られてしまうと Cohen は述べている。そして，もっと大きな標本をとること，そしてできれば検定力分析を行って標本の大きさを合理的に決定することを推奨している。

Cohen は上記の論文に加え，1969年には検定力分析のためのハンドブックを出版し，研究者が自分の目的に合った表を引くだけで，任意の検定力を確保するのに必要な標本の大きさを知ることができるようにした。

それからちょうど20年後に，Sedlmeier & Gigerenzer (1989) は，Cohen のこれらの著作が心理学研究に及ぼした効果の評価を試みた。彼らは1984年発行分の *Journal of Abnormal Psychology* 誌に掲載された論文のうち統計的検定を適用している54編について，Cohen (1962) と同様な検定力分析を行った。その結果，これらの研究における平均的な検定力は，1960年当時よりもさらに低下していることが明らかになった。その原因として彼らは，近年の統計学の発展によって，平均値差に関する多重比較や多変量分散分析といった，第1種の誤りに対してより防衛的な(つまり誤って有意になる可能性を抑える)方法が普及してきたことを挙げている。第1種の誤りと第2種の誤りは，一方の可能性を抑えれば他方の可能性が高まるという関係にあるから，これらの方法の適用は検定力を下げる方向に働くのである。

このような悲観的な結果に対し，Cohen (1992) は，検定力分析が普及しないのは，自分自身の論文や著書を含め，関連する文献が心理学者にとって十分理解できるような内容になっていないためと考え，さらに工夫して単純化した表を発表している。

一方，Sedlmeier & Gigerenzer (1989) は「検定力は本当に必要だろうか」という基本的な問題に立ち返って議論している。そして，現状のように特定の有意水準を遵守し，帰無仮説の棄却・採択といった決定を行い続けるのであれば，検定力の考慮なしに検定を行うのはやはり問題であるとしている。しかし，品質管理で行われるような yes-no 型の決定が心理学研究において適切なものか否かはさらに検討の余地があるとしている。

ところで，検定力は母集団における効果の大きさに依存するため，検定力分析によって標本の大きさを決めるには，「どの程度の大きさの効果なら検出に値するか」，あるいは逆に「どの程度の大きさの効果なら検出しても意味がないか」といった観点から，対立仮説を設定する必要がある。この設定を行うには，効果の大きさの指標に関する記述統計的理解，たとえば，相関係数が .3というのはどの程度の関連の強さなのかとか，効果量が0.5というのはどの程度の分布の重複を意味するのか，といったような理解が最低限必要になる。相関係数から導かれる同順率や，効果量から導かれる優越率は，そうした理

解のために利用することのできる指標の例である（南風原・芝，1987）。

なお，Cohen流の検定力分析は，統計的検定の種類ごとに別々の表を用意するものであるが，検定の種類によらず共通の表を利用する方式も開発されている（Kraemer & Thiemann, 1987）。芝・南風原（1990）は，この方式を紹介するとともに，測定の信頼性を考慮した標本の大きさの決定の問題も論じている。

(2) 実用的有意性の検討

統計的有意性が実用的（または臨床的）有意性とは異なるものであることは明らかである。母集団において実用的に重要な意味をもつ効果が存在しても，検定力が低ければ統計的有意性は得られないし（false negative），逆に検定力が非常に高ければ，実用的には意味のない微小な効果しか存在しなくても，統計的有意性が得られる（false positive）。検定力分析は，実用的に意味のある程度の効果を母集団に想定したとき，それを十分な検定力で検出するための手続である。したがって，検定力分析においてすでに実用的有意性が考慮されていると言うことができるが，ここでは，実用的有意性を問題にしたその他の方法を紹介する。

1つは，帰無仮説を「等平均」や「無相関」といった点仮説とせず，実用的に意味がないと判断される区間からなる仮説とする方法である（Binder, 1963 ; Hodges & Lehmann, 1954）。従来の帰無仮説であれば，仮にそれが棄却されても実用的には意味のない結果かもしれない。これに対し，上記のような区間からなる帰無仮説が棄却されれば，実用的に意味のある効果の存在を主張できるという訳である。これはfalse positiveに対処するための工夫である。

2つめの方法は，同様な発想を，帰無仮説の“立証”を目的とした研究に適用したものであり，等価性検定（equivalency testing）と呼ばれる（Rogers, Howard, & Vessey, 1993）。たとえば，従来の測定手段より簡便な測定手段を用いても実用的には差異がないということを，これら2つの条件間の平均値の“等価性”によって主張したいとき，「実用上，差異がない」と言える等価区間をまず設定する。そして，その等価区間の上限を超える範囲を帰無仮説とした片側検定と，下限を下回る範囲を帰無仮説とした片側検定を行う。そして，そのいずれの帰無仮説も棄却されたとき，実用的に差異がないことを主張するという方法である。これは通常の「等平均」帰無仮説の検定で生じうるfalse negativeに対処するための工夫と言える。

(3) 効果の大きさの推定

これまで述べてきた方法は，効果の大きさの意味付けを，検定を遂行する前に行う点で共通している。これに対し，母集団における効果の大きさをまず推定し，その結果を実質的な観点から解釈しようというのが，ここで述べるアプローチである。

効果の大きさの指標としては，ここまでに触れたものの他に相関比や重相関係数などがあり，さらに従属変数が複数ある場合の多変量版も考案されている（Tatsuoka, 1993）。これらの指標によって，標本のデータに見られる効果の大きさ，関連の強さといったものを記述するのが最初のステップである。次に，これらの指標が標本抽出に伴って確率的に変動する統計量であることから，その変動の大きさを評価することが必要となる。その評価の方法としては，信頼区間の算出と標準誤差の推定が代表的なものである。

標本データに基づいて算出される信頼区間は，それ自体が確率的に変動するものであり，その確率的に変動する区間が母集団値（母数）を含む確率が十分高く（たとえば .95）なるようにしたものである。この確率（信頼係数という）も，統計的検定の場合と同じく，母数を条件としたときのデータの変動に関する条件付き確率であり，1つの標本から算出された区間を固定したときに，母数がその区間に存在する確率を問題にするものではない。実際，信頼区間の理論は（Neyman-Pearson流の）検定理論の一部であり，有意水準 α の両側検定で棄却されない母数の値の集合が，信頼係数 $1-\alpha$ の信頼区間を与える。たとえば，母集団相関係数の95%信頼区間がゼロを含むなら，母集団相関係数がゼロであるという仮説は5%水準の両側検定で棄却されないことになる。

同一の信頼係数に対する信頼区間は，標本が大きいほど狭くなる。したがって，その区間を一定の幅以内に抑えるように標本の大きさを決定することも可能になる。母集団比率などの推定に関しては手計算で必要な標本の大きさを求めることができるが，母集団相関係数などの推定の場合は，検定力分析の場合と同様に特別に用意された数表が必要となる（南風原，1986）。Kupper & Hafner（1989）は，信頼区間が一定の幅以内に収まる確率が所与の値を超えるように標本の大きさを決定する方法について論じている。

信頼係数 $1-\alpha$ の決定は，有意水準の決定と同様に研究者に委ねられている。5%水準の両側検定に対応した95%信頼区間がよく用いられるが，たとえば80%ぐらいの信頼係数で十分と考える研究者がいてもおかしくない。こうした事情を考えると，特定の信頼係数に依存した区間ではなく，標本統計量としての効果の大きさの指標の

標本分布自体に注目し，その標準偏差，すなわち指標の標準誤差を推定して報告するほうが有用かもしれない。その場合，標本の大きさの決定も，標準誤差を基準にして行うことができる。(日本心理学会の「執筆・投稿の手びき」(1991年版)にも平均値等に標準誤差を付すこと，そしてそれを図示することが推奨されている。ただし，標準誤差を範囲や標準偏差などの記述的指標と並べて「散布度」と総称しているのは問題がある。少なくとも，図示されたものが標準誤差なのか標準偏差なのかの区別は明確にされなければならない。)

効果の大きさが推定できれば，次の課題はその結果を解釈することである。効果の大きさの解釈については，(1)で述べたような，指標の値の記述統計的な理解の問題の他に，少なくとも2つの問題がある。その1つは，たとえば相関係数の値が集団の等質性や測度の信頼性の影響を受けるように，効果の大きさの指標が，研究の目的とは直接関係のない種々のデザイン要因(Rubin, 1992)の影響を受けることである。これに対しては，適切な補正をしたり，回帰係数などデザイン要因の影響を受けにくい指標 (Oakes, 1986) を使用したりすることによって，ある程度対処できる。

もう1つの，より根本的な問題は，効果の大きさの実質的な意味付けの難しさである。知能偏差値の尺度での変化の大きさとか，特定の認知的課題への正答率の差異とか，適性検査得点の予測的妥当性とかであれば，こうした意味付けは比較的容易である。この場合は，研究で用いられている変数がそれ自体として研究者の関心の対象であり，その変数の値が彼らにとって直接的な意味をもっているからである。このような，いわば応用的な研究に対し，理論的な仮説を検証することを目指した研究の多くは，その検証のための，いわば“その場限りの道具”として様々な測度を便宜的に利用する。この場合，それらの測度における値が直接的な意味をもっているわけではなく，そのため，その測度に関して得られる効果の大きさの評価も困難となる (Chow, 1988)。

効果の大きさの評価の容易さや，効果の大きさを推定すること自体の有用性が，研究のタイプによって異なるというChow (1988) の指摘は，研究の目的ごとに適切な統計解析の方法が異なるという当然の，しかしややもすると忘れられがちな事実に目を向けさせるものである。しかしながら，彼の「理論研究では効果の大きさは全く無意味である」という主張には賛成できない。研究の道具として便宜的に導入された測度に関して，効果の大きさの細かな差異を問題にすることは確かに意味がないだろうが，その一方で，たとえば相関係数の値が .2という場合と .8という場合では，結果の意味が同じとは言えないだろう。それを，単にどちらも統計的に有意である(あるいは有意でない)として同等に扱うのは，潜在的に有用な情報を切り捨てることになる。

4. 討論

(1) 標本の無作為性と結果の解釈

統計的検定の誤用論の中心テーマの1つは，研究の対象となる標本が特定の母集団から無作為に抽出されているかということである (Bakan, 1966 ; 長谷川, 1994a ; Morrison & Henkel, 1970 ; 橘, 1986)。統計的検定（および推定）において確率的な議論を可能にしているのは，標本抽出過程の確率的性質であるが，実際の研究においては，無作為抽出どころか母集団の定義さえ曖昧であることが珍しくない。もちろん，知能テスト等の標準化においては，明確に規定された母集団から代表的な標本を抽出する努力がなされているし，大規模な社会調査研究では，調査機関などの協力を得て標本の無作為抽出がなされている。しかし，これらは少数の例外である。(興味深いことに，これらの例外においては統計的検定はほとんど用いられない。)

橘 (1986) は，母集団からの無作為標本でなくても，異なる実験条件への被験者の割り付けが無作為に行われれば統計的推論が可能になることを指摘している。しかし，その統計的推論は実際に集められた被験者集団を固定して，その中での無作為な割り付けに伴う確率的な変動のみを考慮したものであり，無作為標本に基づく母集団特性に関する推論とは異質なものである。無作為な割り付けは母集団への推論というより，むしろ因果関係に関する推論の妥当性，すなわち内的妥当性 (Cook & Campbell, 1979) を高めるための操作と言えよう。

それでは，ほとんどの研究で無作為抽出が行われていないという現実と，ほとんどの研究で無作為抽出を前提とした検定が行われているという現実をどう考えたらよいのだろうか。

例として，ある講義を聴講している大学生男女50人ずつにある検査を実施し，男女間の平均得点を比較するというケースを考えてみる。この場合，用いられた被験者集団は明らかに無作為に抽出された標本ではなく，また，性差の検討という問題の性質上，被験者の無作為な割り付けは不可能である。この例において，男子の平均値が女子の平均値より高く，効果量 $d(=(m_1-m_2)/s)$ が0.33となったとしよう。

一方で，同一の正規分布に従う乱数を発生させる装置を用いて，50個ずつの乱数を2群発生させ，その2群間で d を計算することを考える。このとき，d が特定の値0.33を超える確率は約5％ ($p \fallingdotseq .05$) となることが，自由度98のティ分布を利用することによって確かめられる。これは数学的事実である。

ここで問題は，便宜的に用いた被験者集団から得られた結果の解釈に際して，いま述べた数学的事実がどのような意味をもちうるかということである。この問に対して，現実と数学的モデルとの間の明らかな不整合を理由に，上記の数学的事実を全くのナンセンスとする立場は当然考えられる。この立場からすれば，心理学における多くの研究論文は，無意味な数値と，それに基づく無意味な解釈の羅列に過ぎなくなる。

これとは逆に許容的な側では，実際の被験者集団を“仮に”ある母集団からの無作為標本であると“みなし”，その母集団の正規性などを仮定した上で，上記の数学的事実を統計的推論に利用しようという立場が考えられる。これはFisher流の考え方であり，多くの研究者の統計的実践を支持する見方と言えよう。

しかし，この“みなし”は大きな問題をはらんでいる。まず，母集団自体が明確に規定されていない場合は，標本からどこに向けての統計的推論なのか分からない。また，母集団は規定されていても標本の抽出が無作為でなければ，上記の数学的事実をそのまま持ち込んで，「この調査を異なる被験者集団を用いて繰り返すとしたら，母集団における効果がゼロのとき，この程度の差が得られる確率は5％である」といった直接的な推論を行うのは無理がある。ここで言う「調査の繰り返し」が確率過程である保証がないからである。

標本抽出の無作為性が満たされていないこと，したがって，それを仮定した数学的モデルに基づく統計的推論に無理があることを認めたとき，そのモデルの中で成立する上記の数学的事実は，どのような意味をもちうるだろうか。私の考えは，それは結果の「記述的解釈の補助」としてなら意味をもちうる，というものである。つまり，上の例で得られた$d=0.33$という値を解釈する際に，特定の数学的モデルの中では上記の数学的事実($p \fallingdotseq .05$)が成り立つということが，補助的情報として利用できる可能性があるということである。もちろん，そのような情報は解釈の補助にならないという立場もありうるが，いずれにしても，上記の数学的事実の利用価値，あるいは，より一般的に，「統計的に有意である」とか「有意でない」とかという情報の意義は，たかだかその程度ではないかということである。(先に述べたように，被験者の無作為な割り付けがなされれば，それなりの推論が可能である。)

無作為標本でない場合の統計的検定(および推定)の意義をこのように限定して考えると，母集団分布の正規性や等分散性といった統計的推論の前提条件のもつ意味も当然変わってくる。そもそも母集団が特定されない状況では，その分布の特性について議論することはできない。また，仮に母集団は特定のものが想定されるとしても，そこからの無作為な(または代表的な)標本が得られなければ，それらの前提条件の正否を検討する基盤がないことになる。もちろん，このような場合でも，形式的には標本データを用いて正規性の検定や等分散性の検定などを遂行することができるが，その結果が具体的に何を意味するかは定かでない。(先の議論をここにも適用するなら，データとして得られた分布の形や分散の違いを記述的に解釈する際の，補助的情報としての意味はもちうる，ということになる。)

ところで，統計的検定を記述的解釈の補助として利用するとすれば，本稿の冒頭で言及した「『教育心理学研究』と『心理学研究』の間の“検定論文”の比率の有意差の検定」も不可能ではなくなる。普通なら，特定の期間に発行された論文について「全数調査」がなされているから検定は意味がないと判断されるだろう。しかし，データとして得られた比率の差の大きさを解釈する際に，「比率の等しい2つのベルヌイ母集団からの無作為標本で，それ以上の比率差が得られる確率はいくらか」という情報を参考にしてはいけない理由はない。かえって，この場合，推測統計学的なモデルが“架空の状況”であることが明確な分，誤解される可能性が低いと言えるかもしれない。

なお，標本の無作為性が成立しないときに統計的検定を適用し，その結果を上記のような限定された意味で解釈する場合でも，確率の条件性や検定力についての認識は必要である。“架空の状況”であれ，その中で何が言えるのか，そしてどういう要因によってそれが影響されるのかを知らなくては，上記のような解釈ですら適切にはできないからである。

(2) 研究結果の一般化と研究計画

統計的検定や推定の適用の際には，標本から母集団への「一般化」ということがよく言われる。標本は母集団の一部であるから，その一部における結果をもとに，母集団全体にかかわる推論をすることが一般化と呼ばれることに問題はない。ただし，その一般化は標本統計量の値から，それに対応する母集団分布の特性値(母数)への一般化であり，たとえば「母集団においても処遇の効果がある」という結論は，「母集団の全成員において効果がある」ということではなく，「効果の見られる成員の割合が50％を超える」とか「処遇群の平均が統制群の平均を上回る」とかを意味するだけである(Bakan, 1966；橘, 1986)。

また，このような意味での一般化も，(1)で述べた標本の無作為性(あるいは代表性)が前提となっている。実験条件への無作為な割り付けが行われるにしても，もともとが便宜的に集められた標本であれば，単にそれを分割す

るだけであるから，より大きな集団への統計的な一般化は原理的に不可能である。

無作為標本でなく，したがって統計的検定や推定が記述的解釈の補助としての意味しかもたないときには，研究結果の一般化の問題は，その標本と一般化のターゲットとなる集団との間の特徴比較に基づいて，非統計学的な形で議論されることになる(橘，1986)。また，一般化のターゲットとなる集団が特定されない場合に，結果の一般性の範囲・限界を実証的に明らかにするには，多様な特徴をもつ標本による追試を重ねていく必要がある。

追試による一般性の検討という考え方は，1つの標本を構成する個々の成員というミクロなレベルに適用することもできる。たとえば，ある処遇の効果を被験者内要因の実験で評価するときに，個々の被験者ごとにどの程度の効果が見られるかにまず注目し，そうした結果の集積("追試"の積み重ね)として集団のデータをとらえるのである。これは，標本統計量から母数へという一般化とは異なり，「特定の個人に観察される現象が，どれだけ多くの人に見られるのか」という，より素朴な意味での一般化(南風原，1994)を問題にする姿勢を反映しており，事例研究(Firestone, 1993；吉村，1989)や$N=1$実験(岩本・川俣，1990；橘，1986)の目指す方向に通じるアプローチである。また，検定で問題にされる「効果の存在」や，推定で問題にされる「効果の大きさ」に対し，「効果の一般性」を問題にするアプローチとも言えよう。(効果の大きさの指標の解釈のために提案された南風原・芝(1987)の確率的な指標のうち，対になったデータの場合の優越率は，被験者内要因の効果の分析において「正の効果が見られた被験者の割合」を表わすものであり，効果の一般性の指標の1つとみなすことができる。)

統計的検定が用いられている研究を，個に注目し，そこから出発して一般性を問題にしていくという観点から見直すと，母数に関する推論(群平均の比較など)が偏重されている印象を受ける。本稿の冒頭で触れたように，分析方法としてまず統計的検定が選ばれ，その分析が可能となるように研究の計画を立てるということになれば，母数指向の研究が多くなるのは避けられない。問題は，研究で明らかにしたい内容，そして目指している一般化のタイプに照らして，また，標本の無作為性という基本的な前提条件を満たすことが困難な現実に照らして，そのようなアプローチが最適かどうかということである。

統計的検定の限界論や批判論に対しては，「それでは検定に代わる分析法を提案せよ」という要求が出ることが多い(Gigerenzer, 1993)。そうした要求に直接応えるような，すなわち，統計的検定を前提として集められたデータに対して，代替的な分析法を考案する努力もさらになされる必要があろう。しかし，それと同時に，統計的検定を前提としない研究計画(奈須，1992；尾見・川野，1994a；佐藤，1992；山田，1986)にも，もっと目を向けるべきではないだろうか。

(3) 統計教育の課題

本稿で取り上げた内容は，現在，私自身が教育心理学専攻の大学院生を対象に行っている統計教育の内容の一部である。大学院における統計教育の実態は様々であろうが，究極的な目標は統計的方法全般に関して主体的で知的な判断のできる力を養うことであろう。そのためには，技法の習得だけを目標とするのではなく，それぞれの方法の基本原理を理解し，研究方法全体の中でそれらの方法を適切に位置づけることに重点をおいた教育が必要である(長谷川，1994a)。そして，その中で，既存の統計的方法に対する相対的・批判的視点を育てていくことが期待される。

しかし，このような目標を達成する上で障害となる要因がいくつか存在する。1つは，学生自身の統計観，ないしは統計学習観である。彼らの多くにとって，既存の統計的方法は批判的にとらえる対象などではなく，単に技法習得を迫るだけのものである。また相対的・批判的視点の重要性を認識したとしても，学位取得や就職といった現実的な課題の前に，研究の中身とは直接関係のない(ように見える)統計的問題に時間とエネルギーを費やし，場合によっては主体的な判断や創意工夫をしたために新たな誤りを引き起こしたり，論文審査者との間でもめたりするより，広く受け入れられている方法をそのまま踏襲しておこうと考えるのも無理はない。

これと関連するもう1つの要因は，若い研究者を指導する立場にある人々(指導教員や論文審査者)の統計観や統計学的知識である。結局のところ，この層の人々の過去の統計的実践(もちろん全部ではないが)が統計誤用論のターゲットにされてきたのであるから，それが無批判に次世代に教え継がれてはならない。

検定力分析の普及に努めてきたCohenは，その多大な努力がなかなか効を奏さない現実に対して，「変化は徐々にしか起こらない」と述懐している(Cohen, 1990)。検定力分析がもっと普及し，その結果として統計的検定がさらに隆盛を極めるというのはまた問題だが，上記のような目標に向けての統計教育の広がりや，その効果についても，「変化は徐々にしか起こらない」ということは言えそうである。

引用文献

安藤洋美　1989　統計学けんか物語―カール・ピアソン一代記　海鳴社

Bakan, D. 1966 The test of significance in psychological research. *Psychological Bulletin*, **66**, 423-437.

Berger, J.O., & Sellke, T. 1987 Testing a point null hypothesis : The irreconcilability of *P* values and evidence. *Journal of the American Statistical Association*, **82**, 112-122. (with discussion, 123-139).

Binder, A. 1963 Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, **70**, 107-115.

Casella, G., & Berger, R.L. 1987 Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, **82**, 106-111. (with discussion, 123-139).

Chow, S.L. 1988 Significance test or effect size ? *Psychological Bulletin*, **103**, 105-110.

Cohen, J. 1962 The statistical power of abnormal-social psychological research : A review. *Journal of Abnormal and Social Psychology*, **65**, 145-153.

Cohen, J. 1969 *Statistical power analysis for the behavioral sciences*. San Diego : Academic Press.

Cohen, J. 1990 Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.

Cohen, J. 1992 A power primer. *Psychological Bulletin*, **112**, 155-159.

Cook, T.D., & Campbell, D.T. 1979 *Quasi-experimentation : Design & analysis issues for field settings*. Chicago : Rand McNally.

Cowles, M. 1989 *Statistics in psychology : An historical perspective*. Hillsdale, NJ : Erlbaum.

Evans, J.St.B.T. 1989 *Bias in human reasoning : Causes and consequences*. Hillsdale, NJ : Erlbaum.

Firestone, W.A. 1993 Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, **22**(4), 16-23.

Fisher, R.A. 1925 *Statistical methods for research workers*. New York : Hafner.

Fisher, R.A. 1935 *The design of experiments*. New York : Hafner.

Fisher, R.A. 1956 *Statistical methods and scientific inference*. New York : Hafner. (渋谷政昭・竹内啓(訳) 1962 統計的方法と科学的推論 岩波書店)

Fisher, R.A. 1990 *Statistical methods, experimental design, and scientific inference*. Oxford : Oxford University Press.

Gigerenzer, G. 1993 The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences : Methodological issues*. Hillsdale, NJ : Erlbaum. Pp.311-339.

Gigerenzer, G., & Murray, D.J. 1987 *Cognition as intuitive statistics*. Hillsdale, NJ : Erlbaum.

南風原朝和 1986 相関係数を用いる研究において被験者数を決めるための簡便な表 教育心理学研究, **34**, 155-158.

南風原朝和 1994 一般化可能性の評価と有意性検定 日本心理学会第58回大会発表論文集, S11.

南風原朝和・芝 祐順 1987 相関係数および平均値差の解釈のための確率的な指標 教育心理学研究, **35**, 259-265.

長谷川芳典 1994a 心理学研究法再考—(1)基礎的統計解析の誤用をなくすための30のチェック項目 岡山大学文学部紀要, **21**, 47-59.

長谷川芳典 1994b 心理学研究における基礎的統計解析の誤用 日本心理学会第58回大会発表論文集, 8.

Hodges, J.L., & Lehmann, E.L. 1954 Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B*, **16**, 261-281.

岩本隆茂・川俣甲子夫 1990 シングル・ケース研究法—新しい実験計画法とその応用 勁草書房

Kraemer, H.C., & Thiemann, S. 1987 *How many subjects ?* Newbury Park, CA : Sage.

Kupper, L.L., & Hafner, K.B. 1989 How appropriate are popular sample size formulas ? *American Statistician*, **43**, 101-105.

Morrison, D.E., & Henkel, R.E.(Eds.) 1970 *The significance test controversy : A reader*. Chicago : Aldine. (内海庫一郎・杉森滉一・木村和範(訳) 1980 統計的検定は有効か—有意性検定論争 梓出版社)

奈須正浩 1992 教育研究と数量的方法—教育におけるリアリティの特質に注目して 教育方法史研究(東京大学教育学部教育方法学研究室), **4**, 225-244.

Neyman, J., & Pearson, E.S. 1933 On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, **231**, 289-337.

日本心理学会 1991 執筆・投稿の手びき 1991年度版 日本心理学会

Oakes, M. 1986 *Statistical inference : A commen-*

tary for the social and behavioural sciences. Chichester : Wiley.

尾見康博・川野健治　1994a　人びとの生活を記述する心理学―もうひとつの方法論をめぐって　東京都立大学心理学研究，**4**，11-18.

尾見康博・川野健治　1994b　心理学における統計手法再考―数字に対する“期待”と“不安”　性格心理学研究，**2**，56-67.

Pollard, P.　1993　How significant is “significance” ? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences : Methodological issues*. Hillsdale, NJ : Erlbaum. Pp.449-460.

Rogers, J.L., Howard, K.I., & Vessey, J.T.　1993　Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, **113**, 553-565.

Rosenthal, R., & Rosnow, R.L.　1984　*Essentials of behavioral research : Methods and data analysis*. New York : McGraw-Hill.

Rubin, D.B.　1992　Meta-analysis : Literature synthesis or effect-size surface estimation ? *Journal of Educational Statistics*, **17**, 363-374.

佐藤郁哉　1992　フィールドワーク―書を持って街へ出よう　新曜社

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

芝　祐順・南風原朝和　1990　行動科学における統計解析法　東京大学出版会

繁桝算男　1985　ベイズ統計入門　東京大学出版会

繁桝算男(企画)　1994　シンポジウム：教育心理学における統計的方法―DOs and DON'Ts　日本教育心理学会第36回総会発表論文集，46-47.

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　医療図書出版社

Tatsuoka, M.M.　1993　Effect size. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences : Methodological issues*. Hillsdale, NJ : Erlbaum. Pp.461-479.

山田洋子　1986　モデル構成をめざす現場心理学の方法論　愛知淑徳短期大学研究紀要，**25**，31-50.

吉村浩一　1989　心理学における事例研究法の役割　心理学評論，**32**，177-196.