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*

長　谷　川　芳　典

1. はじめに

本稿は，研究論文や学会発表のさいの自己点検項目として，あるいは指導上の指針として活用されることを目的として，心理学研究における基礎的な統計解析の誤用を克服するための30のチェック項目を提起するものである．

統計の誤用は，心理学の論文ではかなりの比率で認められるものである．私自身も過去に多くの誤りをおかしてきた．“赤信号みんなで渡ればこわくない”という言葉がある．もし心理学者たちの多くが，この誤りを繰り返し許容していたとするなら事は重大である．統計の誤用は英語の誤用とは異なる．英語で書かれた論文に文法上の誤りや稚拙な表現があったとしても，内容が誤って伝えられない限りは大目にみてよろしかろう．しかし，統計の誤用は，結論やその後の研究の方向に致命的な影響を及ぼす．心理学の研究を正しく発展させるためには，心理学者自らが日常的に自己点検をしていかなければならない．

1993年度の心理学会では個人の研究発表は870件にのぼった．しかし統計的誤用の問題を放置したままで，今後この数が2000，3000と増えたとしても心理学の発展にはまったくつながらない．むしろ，870の件数が400に減ったとしても，統計的方法を正しく用いた良質の研究が100から200に増えるならばそのほうがよほど価値がある．

本稿は，けっして他の心理学者を攻撃中傷するために意図されたものではない．心理学者が自ら使用している統計解析の方法について自己点検を行ない，誤用についての建設的な相互批判がなされることを目ざして，問題提起を行なうものである．

なお具体的な誤用の実例や誤用頻度については，1994年10月開催予定の日本心理学会第58回大会ワークショップ“心理学研究の自己評価　(1)基礎的統計解析の誤用と対策（企画者：長谷川芳典，基調講演：橘敏明氏，指定討論者：桑田繁氏)”における討論をふまえ，参加者と協議のうえで別のかたちで報告をまとめていきたいと考えている．

＊本稿は岡山心理学会第41回大会（1993年12月11日）の抄録原稿を加筆修正したものである．

2. 誤用の分類

2.1. データを集める前の誤用

2.1.1. 母集団の規定

推測統計の基本は標本から母集団を推測することにある［補注1］．母集団が何かということをあいまいにした解析は意味をなさない．

同じ標本を検討する場合でも，研究の目的によって母集団の規定は変わってくる．たとえば，次のような問題（仮想データ）を考えてみよう．

"ある大学に10000人の学生がいたとする．毎年，このうちの300人を無作為抽出して女子学生の人数を調べたところ，10年前は40%だったものが本年度は50%になっていた．女子は10年前に比べて増加したと言えるだろうか."

この問題は，ほんらいは選挙の得票率の予測と同様であり，300人のサンプルから10000人という母集団における女子の比率を推定することにある．したがって名簿などから全数調査が可能となった段階で，女子が増えたかどうかははっきりと断定できる．もし10年前の女子学生が10000人中4000人であり本年度が4001人であったとしたら，女子学生は事実として増加したことになり，もはや何の推測も必要とならない．これに対して，"日本では女子の大学進学者は増加傾向にある"という一般的仮説を検討しようとしている者にとっては，母集団はもっと大きなものになる．また4000人が4001人に増えたという事実がわかったとしても，自分の仮説の証拠にはならないだろう（もっとも1つの大学の標本だけからこのような仮説を検証すること自体，問題であろうが）．

【チェック項目1】母集団は何を想定しているのか？

2.1.2. 無作為抽出

橘(1986, p.33)も指摘しているように，母集団の要素すべてがリストアップされたなかから標本を無作為に抽出するということは，現実の研究ではほとんど不可能に近い．しかし，標本の取り方に偏りがでないよう最大限に努力する必要はあるだろう．

卒論実験などでは，心理学受講生やその知り合い，サークル仲間などを被験者に依頼する場合が多いが，それらの人々は無作為に抽出された標本にはなっていない．

卒論研究などでは，当初の実験計画で2群を比較する予定であったものが，その後考え直して，もう1条件を設定して3群で比較するというように計画を変更する場合がある．そのさい，2群のデータをとり終えたあとで3番目の群の被験者集めをするのは甚だ問題が多い．特に友人などを頼って被験者集めをしている場合は，あとから集める被験者ほど，知らない人や実験者とは違う学部の人に依頼する頻度が高くなる．また，被験者を集める時期が異なれば，試験勉強に追われている被験者ばかりに偏る可能性や，実験室外を含めた気温のちがいが実験の遂行に思いもよらない影響を及ぼす可能性もある．

【チェック項目２】標本の無作為抽出にどこまで努力したか？
【チェック項目３】標本は群間で等質になっているか？

2.1.3. 検証実験と反証実験の区別

心理学の実験には，仮説を検証するための実験のほか，反例を示すことで特定理論や一般常識観念をくつがえすことをねらった実験もある．研究の目的が検証にあるのか反証にあるのかによって統計解析の仕方もかわってくる．

たとえば，チンパンジーの言語学習の実験を考えてみよう．種としてのチンパンジーが特定の文法規則を理解できることを示そうとするならば，チンパンジーという母集団から無作為に抽出した被験体に言語訓練を施して検証する必要があるだろう．いっぽう，その文法規則は人間しか理解できないということが理論や一般常識観念として背景にあるのならば，賢いチンパンジー１頭を被験体として，“ここまでできる”という反例を１つでも示せば事足りる．

私はかつて，幼児１名を被験者として，日本語の単熟語はひらがな表記より漢字表記のほうが速く読めるというような発表をしたことがある（長谷川，1988；1989b）．これらは反例を示す実験であり，現行のひらがな先行の文字教育，あるいは“ひらがなのほうが覚えやすい”というような一般的固定観念に対して“むしろ漢字のほうが覚えやすい”という一例を提供することが目的であった．

私はまた，“「血液型と性格」についての非科学的俗説を否定する”というような発表をしたことがある（長谷川，1985；1987；1989a）．これらの発表も「血液型人間学」などと呼ばれる俗説的固定観念に対して反例を示すことを目的としたものであった．つまり，もし血液型によって性格や行動特性に顕著な差があるならば，公開講座受講生とか心理学概論受講生といった任意に選んだサンプルの中でも有意な差が見られて当然であろう．しかし，結果は，どの血液型においても同じような比率で多様な性格類型が見られたのであった．この場合，公開講座受講生とか心理学概論受講生というサンプルは日本人あるいは人類全体から無作為抽出したサンプルではなく，人数も調査により94～752名で比較的少数になっている．これらのデータからは“血液型と性格は関係ない”という一般的結論を引き出すことはできない．あくまで“血液型と性格は関係があるかもしれません．しかし，少なくとも職業適性，相性，対人行動などの面では，血液型を知っても何の役にも立ちません”という反例を示したものであったことを強調しておきたい（長谷川，1994参照）．

【チェック項目４】検証のための実験か，反例を示すための実験か？

2.1.4. “探索的？”な“検定”

100個の比較項目を順々にt検定していけば，あるいは100群それぞれの内部の構成比率をχ^2検定していけば，ほんらい偶然的な片寄りでありながら「有意」と判定される項目（群）が５個くらい見つかるのは当然であろう（$p<.05$だから５個くらいと言っているわけではない．2.3.1参照）．「有意差」のあった5項目（群）だけをとりあげて独立変数の効果であるように解釈するのは誇大広告と

同類であり，ちょうど民間療法の宣伝のチラシで治った事例ばかりを紹介するようなものである．

比較可能な項目(群)が100種類あったならば，そのうちなぜ5項目(群)をとりあげたのかを事前に明記しておく必要がある．もしくは，事後に，残りの95項目(群)がなぜ有意でなかったのかを明確に説明しなければならない．

このトリックを意図的に利用したものとして血液型人間「学」をあげることができる．彼らは，しばしば多数のなかから血液型比率に見かけ上の偏りがある職業だけを抜き出し，こじつけ的な説明を行なっている．もし10種類の職業をサンプルとして“職業によって血液型比率に差がある”ことを統計的に実証しようとするならば，職業の種類そのものが無作為に抽出されていなければならない．無作為法ではなく，たとえばスポーツの能力に秀でた集団の血液型の比率を問題にする場合でも，比率に「偏り」が出た種目だけを事後解釈するのではなく，未調査の種目について血液型の比率を予測するなど，検証可能なかたちで仮説を提示する義務があるだろう．

心理学の実験研究においても同様の誤用が起こる可能性がある．たとえば，40匹のネズミを2群に分け，“縦縞群”は縦縞模様の箱の中で，“横縞群”は横縞模様の箱の中でそれぞれ集団飼育し，ストレスの大きさを比較したとしよう．そしてその結果，当初予想していた体重，活動性，毛のハゲ具合には両群に有意な差が認められなかったとする．しかし，“横縞のほうがストレスが大きい”という自説に固執しているとこの結果には満足できない．そこで，新たに両群のシッポの長さ，ヒゲの数，糞の成分など，ありとあらゆる指標を比較検定することになる．そしてたまたまヒゲの数に有意な差があれば，“縞模様の違いはヒゲの数に有意な影響を及ぼした”として自説の正しさを力説することになるだろう．しかし科学的態度を守るならば，有意差が生じた比較項目だけを論じるのではなく，ヒゲ以外の比較項目ではなぜ有意差が生じなかったのか，ヒゲだけで差が生じたのはどのような生理的原因によるのかということをフェアに考察する必要があるだろう．

【チェック項目5】有意差の出た比較項目だけを過大にとりあげていないか？

2.2. データを分析するさいの誤用

2.2.1. 何でもかんでも平均

データを集めると何でもかんでも平均をとり，グループ間で比較しようとする傾向がある．こうした“平均値信仰”そのものの弊害については別の機会にも指摘したことがあるので(たとえば，長谷川，1993，p.50-52)，ここでは誤用に関する話題だけを取り上げることにしよう．

そもそも，平均値（算術平均値）とは，測定対象が間隔尺度以上である場合に限って意味がある代表値である．

“全くあてはまらない”，“当てはまる”などの形容詞をつけて5件法や7件法で回答させるようなデータは，ほんらい順序尺度であるので，平均値を算出することはできない．しかし，現実には，何のためらいもなくそれらを得点化し，平均を求める傾向がある．たんに“高度な統計解析が可能になるから”という理由や“みんながしているから”という理由だけで，順序尺度を間隔尺度にすり替えてしまってよいものだろうか．

あるデータの物理的な単位が間隔尺度以上であることと，それが比較指標として間隔尺度であることとは別問題である．たとえば，エルニーニョ現象の影響を調べるため，ある地点の①地上気温，②上空の気温，③井戸水の温度を測り，平年の値と比較したとしよう．この場合，温度（摂氏）という単位自体は間隔尺度であるが，①，②，③を平均して比較したところで何の意味もない．

同様の理由で，筋電位の積分値のようなデータを平均することにも疑義がある．たとえば，あるリラクセーション訓練によって，Ａさんの値が30000から25000に，Ｂさんの値が20000から15000に減ったとしよう．この場合，両者の減少値5000が生理学的に同じ意味をもっていなければ，２人の平均値を求めたりそれらをもとにｔ検定や分散分析をすることは意味がない．

【チェック項目６】データは名義尺度か，順序尺度か，間隔尺度か，比率尺度か？
【チェック項目７】そのデータから算術平均を求めることは妥当であるか？

2.2.2. 標本分散と不偏分散

関数機能付き電卓や統計パッケージを用いると，数値を投入しただけで自動的に分散が出力される．そのさい，求められた値が，平均値からの偏差の自乗の合計を標本数ｎで割った値（記述統計レベルでの分散，ここでは標本分散と呼ぶ）であるのか，それともｎ－１で割った不偏分散であるのか，はっきりと区別しておく必要がある．特にnが小さいときには，これらを混同すると算出値にズレが生じることになる．

なお，しばしば誤解されているが，不偏分散の正の平方根をとっても母数σの不偏推定量にはならない．このあたりの議論については，岩原 (1965, p.59-60)，岡田(1966, p.111-113)などを参照されたい．

【チェック項目８】標本分散か，不偏分散か明記されているか？

2.2.3. 何でもかんでも相関係数

相関係数に関する誤用としては，順序尺度のデータからピアソンの相関係数を算出してしまう誤りが多い．順序尺度ならばスピアマンやケンドールの順位相関係数を求めるべきであろう．論文や発表では，ただ“○○について相関係数を算出した”と書くのではなく，どういう相関係数をどういう理由で用いたのかを明記する必要がある．

【チェック項目９】どういう種類の相関係数をどういう理由で用いたのか？

2.2.4. 何でもかんでもｔ検定

ｔ検定に関する誤用を３つほどあげれば，まず，ｔ検定の前提条件（母集団は正規分布，等分散，標本は無作為抽出）を無視している場合．つぎに，３群以上の比較において多重比較をすべきところをあらゆる２群間の組み合わせについてｔ検定をするという誤りがある．また，もとのデータが

連続変量（たとえば年齢とか尺度得点）であるのをわざわざグループ分けして群間でｔ検定を行なうような誤りも指摘されている（橘，1986，p.107）．いずれも統計学の勉強不足に起因するものと思われる．

【チェック項目10】ｔ検定の前提を満たしているか？
【チェック項目11】多重比較すべきところにｔ検定を用いていないか？
【チェック項目12】ｔ検定にかけるために無意味なグループ分けをしていないか？
【チェック項目13】ｔ検定と分散分析しか知らないから使っているのではないか？

2.2.5. 片側検定か両側検定か

両側，片側のどちらを用いるかは，ほんらいデータを集める前に決めておくべきことである．両側検定をするつもりだったが，データを集めた後に平均がＡ＞Ｂだったから片側検定に変更するなどというのは，検定の大原則に反する．

片側検定（例：Ａ≦Ｂを棄却してＡ＞Ｂを見出す）は，本質的にＡ＜Ｂがありえない場合［補注2］，あるいはＡ＜Ｂを考慮に入れる必要がない場合［補注3］に用いられる．これら以外の場合は，原則として両側検定が推奨される［補注4］．

両側検定の場合は“ＡとＢの差は有意であった”とは言えるが，“ＡよりＢのほうが有意に大であった”とは言えない．Ａ＞Ｂと結論することは実際的には問題がないが（近藤・安藤，1967，p.16），これは検定の結果ではなくて，信頼限界に基づく推定の結果であることを理解しておかなければならない．

【チェック項目14】両側検定か片側検定か，明記されているか？
【チェック項目15】片側検定とする根拠が明示されているか？

2.2.6. 標本の独立性

標本の独立性は多くの統計解析の前提になっている．この前提条件を忘れて機械的にデータを公式に投入すると，過度に有意差が出やすくなる．この危険のあるケースとして，①互いに影響を及ぼし合うような同一集団内の構成員を被験者（体）とする場合，②同一個体内で一定時間内に変化する観測値をデータとする場合，の２点をあげておこう．

①の例として，予防注射を実施した学級と実施しなかった学級のインフルエンザの発生率を比較したとする．インフルエンザに感染性がある以上，ここでは標本の独立性は保たれていない．（橘，1986，p.44参照）．

動物の集団飼育実験において異なる環境が生育に与える影響を検討する場合も同様である．病気の発生，摂食量，飼育環境下での活動量などは他個体の状態に依存しやすい．したがって，これらの影響を受けるであろう死亡数，体重，活動量などを従属変数として扱う場合は，慎重な配慮が必要である．たとえば，2.1.4.でかかげたネズミの集団飼育の実験で，１年以内に横縞群の18匹，縦

縞群の2匹が死亡したとしよう．この場合，単純に検定をすればC.R.＝5.06だから有意であるというような結果が導かれるが，横縞群の1匹がたまたま病気になり他の17匹にも感染して死亡した可能性，あるいは横縞群の中にたまたま“多動”ネズミがいて他のネズミの睡眠を妨げストレス死をもたらした可能性なども考慮しなければならない．もしこれらが起こる可能性があるのなら，20匹中18匹とか2匹という数は独立標本の数とは言えず，不当に大きなC.R.が算出される恐れがある．

次に②については，データの系列依存が大きな問題となる．たとえば，ある薬に体温を上昇させる影響があるかどうかを調べるため，服用後24時間の体温を1時間おきに測定したとする．この場合，形式的には24個のデータが得られるが，体温の変化は時系列上の状態に依存して変化するものであるから，これらは独立した観測値とは言えない．

このほか，データの系列依存性がもたらす問題については，バーロー・ハーセン(1993, p.196-199)をあわせて参照されたい．

【チェック項目16】標本の独立性は保証されているか？
【チェック項目17】集団内で互いに影響を及ぼし合うようなデータではないか？
【チェック項目18】系列依存のおこりやすいデータではないか？

2.2.7. X^2検定における作為的な事後分類

橘(1986, p.30)はこの問題について“カイ2乗検定に用いる分割表のカテゴリー（すなわち群）の取り扱いでしばしばみられる誤りは，結果が出てからカテゴリーを作ったり，結果が出てからカテゴリーをプールしたりすることである”と指摘している．事後的な任意の分類を許せば，カテゴリー分けを都合よく変えることによって，同じデータから“有意差あり”という結果も“有意差なし”という結果も思いのままに引き出すことができるようになる．分割表におけるカテゴリー分けは，正当な理由のもとに事前に行なわれなければならない．

【チェック項目19】カテゴリー分けは正当な理由のもとで行なわれているか？

2.2.8. 数量データを無理やりカテゴリー分けする

独立変数や従属変数が連続変量であるにもかかわらず，それらを“上，中，下”群のように分けて分析することをいう．

たとえば，調査の段階で具体的な年齢を質問しているにもかかわらず，分析の段階では“30歳以上”群と“30歳未満”群に分けて各群の得点の差を検定したり頻度のχ^2検定をしたうえで，“○○については年齢による差が認められた”というように結論を下している発表を聞いたことがある．これとは別に，大学生に不安検査を行ない，不安得点に応じて“高不安群”，“中不安群”，“低不安群”に分けて分析をしている発表を聞いたことがある．これらの例は少なくとも3つの問題点を含んでいる．1つは2.2.7にも述べたように，事後的に作為的なカテゴリー分けが行なわれる可能性

があること，第2に，連続変量がもつ種々の貴重な情報をわざわざ捨てていることである．第3は，むしろ結果の解釈にかかわることであるが，χ^2検定の結果だけから，もともとの連続変量と従属変数とのあいだに量的な相関があるというように判断を下す危険がある点である．たとえば上の例で“年齢による差”とは単に30歳以上と未満の差を示しているにすぎないが，これがいつの間にか“年齢が高いほど○○の傾向がある”という考察にすりかわってしまう恐れに注意しなければならない．なお，不安検査の例についてはもうひとつ，“高不安群”とか“低不安群”とか言ってもあくまで健常者の範囲の中での高低であって，病的な不安や異常行動の解明にはつながらない点に留意しておく必要がある．

橘(1986)は，“自分の知っている検定法にむりやり押し込める(p.107)”ことが，こうした傾向の一因であろうと指摘している．“何でもかんでも平均”，“何でもかんでもｔ検定”，“何でもかんでも相関係数”，あるいは今回はとりあげていないが“何でもかんでも分散分析”，“何でもかんでもχ^2検定”という傾向も，すべて統計解析についての勉強不足に起因することが多いように思う．

【チェック項目20】カテゴリー分けで連続変量のもつ貴重な情報が失われていないか？
【チェック項目21】正常範囲内の相対比較から異常現象を拡大解釈していないか？

2.3. 結果の解釈をめぐる誤用

2.3.1. Ｐ値の意味

統計的検定で得られるＰ値についてはまず，橘(1986)が指摘した以下の点に留意する必要がある(p.63～73ほか)．

・Ｐや１－Ｐは対立仮説の正誤を示す確率ではない．
・Ｐや１－Ｐは帰無仮説の正誤を示す確率ではない．
・Ｐ値は結果の再現性を反映しているわけではない．
・Ｐは，もし帰無仮説が正しいとすると得られた標本の差が生じる確率はＰ以下であるという意味である．
・Ｐは，帰無仮説が真のもとでの“観測値の出現率”を計算しているにすぎない．

Ｐの解釈は学派によって異なると言われるが，少なくとも両群の差の程度を表すものではない．ある口頭発表では，“Ａ条件のもとでは両群の差は５％で有意でありＢ条件のもとでは１％で有意であるからＢ条件のほうが差が大きい”などと主張していたが，これはまったくの誤りである．また別の発表では，“ｔの値があと0.2だけ大きかったら有意であった”とか“ｐ＜.07なので有意ではないが有意な傾向があった”というような発言が聞かれたが，これらもＰの意味をわきまえた解釈とは言いがたい．

このほか，相関係数が有意であるとｒ＝1.00であるかのように解釈されてしまう誤用例が橘(1986，p.101)によって指摘されている．

【チェック項目22】有意水準の違いを“差の大きさ”と混同していないか？
【チェック項目23】p＞.05という結果が出た場合に誤った解釈をしていないか？
【チェック項目24】相関係数が有意であることをｒ＝1.00と混同していないか？

2.3.2. 有意差と傾向

χ^2検定による度数の検定では，“各組の標本は度数比が一様な母集団からのランダム標本である”というのが帰無仮説になる．ところが，これが棄却された場合にあたかも“傾向”が実証されたかのように結論してしまう誤用がある．

一例としてランダムに選ばれた社会人男性を年齢別に３群に分け，Ａ，Ｂ，Ｃ，３政党のいずれを支持するか回答を求め，年齢によって支持政党が異なるかどうかをχ^2検定したとしよう．もし有意であった場合の結論はあくまで“年齢によって支持政党は異なる”ということにとどまるはずであるが，“年齢が高いほどＡ政党を支持している”というような“傾向を示唆する”ような結論を出してしまうのである（2.2.8.を合わせて参照）．

別の例として，何人かの被験者に学習訓練を5試行反復し試行間で成績に有意な差があるのかをFriedman検定したとしよう．この場合，有意な差とはあくまで試行間の差を意味しているはずであるが，“試行を反復するにつれて成績が向上した”といった“傾向”に言及した結論を出してしまう誤用が見られる．

【チェック項目25】連続変量の傾向をX^2検定だけから過大に解釈していないか？

2.3.3. 解析結果の過大な一般化

統計解析が常に最適の科学的解決をもたらすかどうかは断言できない．橘(1986, p.88-97)が指摘しているように，特に有意性検定は，それ自体さまざまな欠点をかかえるものである．

しかし，いったん統計解析を行なうと決めた以上は，解析結果から客観的に導かれる帰結と，研究者が自らの創造力に基づいて行なう推理や解釈とははっきりと区別して記述されなければならない．統計解析の誤用に基づく結果の過大な一般化は厳につつしむべきである．

過大な一般化の例として，まず記述統計と推測統計との混同をあげることにしよう．たとえば，サンプルの相関係数が0.9であったからといって母集団の相関係数が0.9であるかのような議論はできない．

全数調査でありながら想定していない「母集団」の特性がいつのまにか議論されることさえある．2.1.1.に述べたこととも関連するが，全数調査で確認された変化は事実そのものを意味する．かりに心理学講座1回生の学生(15人)の中で灰色の好きな人数が３年前の５人から10人に増えたとしよう．この変化は事実である．いかなる統計的推測も必要としない事実である．しかしこのことは“世の中が不況になったために灰色が好まれるようになった”という主張の根拠にはならない．増えたという事実は，単なる偶然的な変動によるものかもしれない．より決定論的な観点からみても，同定できないほどの多数の原因が増加をもたらしたのかもしれないのである．繰り返し言えば，“不

況の影響”というのは事実に基づいた推論ではない．（不況であるほど灰色の好みが増すという別の統計的根拠が示されない限りは）単なる事後解釈の１つにすぎないのである．

調査対象の内的妥当性（標本「内部」における有意な傾向）が外的妥当性（母集団全体の傾向）と混同される場合もある．たとえば，m匹のネズミに走路を走らせ試行とともに走行時間が短くなっているのかをMテストで傾向検定したとしよう．この場合の結論はm匹のネズミだけにあてはまる傾向である（岩原，1964，p.236-238参照）．

【チェック項目26】全数調査から標本調査的な結論に飛躍していないか？

【チェック項目27】全数調査で見られる「変化」や「ちがい」を１つの「要因」でむりやり説明しようとしていないか？

【チェック項目28】検定結果の内的妥当性と外的妥当性をきっちりと区別しているか？

2.3.4. 観察研究と実験研究

解析結果の解釈にあたって最も重要な点は，無作為な割りつけがなされているかどうかにある．無作為な割りつけは無作為抽出とはまったく別の問題であり，各被験体が各処理群に割りつけられる可能性を等しくするような手続のことをいう．橘(1986)は，独立変数が操作できるかどうかは無作為な割りつけができるか否かにかかっており，このことこそが実験研究と観察研究とを分ける決め手であると述べている(p.35-37)．

観察研究では，未知の独立変数の効果を無作為な割りつけによって取り除くことができないので，かりに群間に有意な差が見られたとしても，研究者が割りつけにあたって想定した要因に基づく差であると断定することはできない．また，因果関係なのか，共通原因がもたらす相関関係であるのかを断定することもできない．

ここで注意しなければならないのは，“実験的”手法を用いても無作為な割りつけができていない場合があるということである．

たとえば，オス40匹，メス40匹のネズミをそれぞれ実験群と統制群にランダムに振り分け，２×２の実験計画でデータを集めたとしよう．この場合，実験群と統制群の振り分けは無作為に割りつけられているので，独立変数が原因となって従属変数の差をもたらしたと結論することができる．しかし，雌雄のデータに有意な差があった場合（あるいは交互作用が有意であった場合）については，雌雄が無作為な割りつけではない以上，それらの原因を無条件に雌雄の違いに帰着させるわけにはいかない［補注５］．

健常児と発達障害児の比較，年齢の差，日本人とブラジル人との比較なども同様の理由ですべて観察研究に分類される．たとえば高齢者と若年者で差が見られたとしても，それらは収入の差であるかもしれないし，家族構成の差であるかもしれない．被験者がいかに無作為に選ばれようとも，そこで得られた群間の有意差の原因は，ただちには独立変数の違いに帰着させるわけにはいかないのである．

【チェック項目29】観察研究の場合，未知の原因を見逃していないだろうか？

3．誤用の原因

現段階では実証的な議論はできないが，誤用の原因として，次の5点が考えられる．

3.1. 基礎的な統計教育の不足

心理学の専門課程は，ほとんどの場合文系の学部に属しているため，確率論を含む基礎的な統計教育が行なわれにくい状況にある．また教官自身が統計を苦手とし，“誤用の宝庫”と化している場合もあるかもしれない．

3.2. 先行研究で用いられた統計的方法を無批判に継承する誤り

さいきんでは，卒論研究でも高度の多変量解析を用いる学生が多い．しかし，統計学以外に学ぶべき内容があまりにも多く，とうてい基本原理を理解する余裕がない．結果的に先行研究が用いた統計的方法を無批判に受け入れたり，統計ソフトのマニュアルに従って機械的にデータをインプットするだけに終わってしまう．

3.3. 統計教育における力点のおき方に問題

統計的技法の高度化に伴なって，技法の習得をあせるあまり，無作為抽出，無作為割りつけ，標本の独立性などに関する説明がおろそかになる．

3.4. 悪しき業績主義

大学教官の公募や科研費採択の審査にあたって業績が客観的に評価されるようになったのは好ましい傾向であるが，論文の質よりも数を競うあまり，“有意差さえ出れば論文が1つ書ける”，“有意差が出ない実験は失敗だ”と考え，基本的前提を無視して何がなんでも有意差を出そうとする風潮が強まっている可能性がある．

3.5. 専門分野の細分化あるいは日本人的ななれあいがもたらす無批判主義

自分のオリジナルな研究領域に没頭するあまり，他分野の研究に対する関心がうすれ，もしくは他分野の研究を評価する力を失ってしまった可能性がある．

また，“他人が努力したことに文句をつけない”ことを美徳とするような日本人的ななれあい主義が，誤用を放置したとも考えられる．

4．誤用の対策

現段階では思いつきの域を出ないが，次の3点を具体化する必要があると思う．

4.1. 技法教育に終始しないような統計教育の充実

いかに技法の習得が遅れようとも，検定の基本的事項については徹底的に教えることが必要である．多くの大学では心理学の講座は文系の学部に所属しているため，確率論についての基礎知識を持たずに入学してくる学生が多い．理系なみに数学を重視した入試をするとか，必修科目としての統計学の授業を増やすなどの工夫がいるだろう．

4.2. 個体内比較法の普及

何がなんでもグループ分けをして，各群の平均値の差を検定することばかりが心理学の研究ではない．むしろ独立変数が個体内の行動の変容に及ぼす効果を系統的に検討したほうが成果が多い場合もある．少なくとも統計教育にあたっては，個体内比較法は個体間比較法と同等に扱われるべきであろう［補注６］．

【チェック項目30】個体間比較法，個体内比較法の長所を生かしているか？

4.3. 相互批判を活発化し少数の良質の研究や独創的な評論を評価する体制を築く

学問は相互の批判があってこそ発展するものである．研究者個々人が細分化された固有の研究分野に没頭するのではなく，心理学の問題全般にわたって相互批判ができるような環境を整える必要があるだろう．

補　　注

［補注１］：正確には，研究対象となる個体の集合はユニバースと呼ばれ，母集団とはそれらの個体の測定値の集合を意味する（森・吉田，1990，p.47参照）．

［補注２］：例えば，小学校２年生の語彙数Ａが１年生の語彙数Ｂより有意に増えているかどうかを検定する場合，語彙数が２年生になって減少することは考えられないのでＡ＜Ｂを考慮する必要はない．

［補注３］：例えば，ある添加物に発癌性があるかどうかを検定するために，その添加物を投与したラットにおける癌の発生率Ａと投与しなかった対照群のラットにおける発生率Ｂを比較したとしよう．この場合，Ａ＜Ｂとなる可能性，つまりその添加物には癌の発生を押さえる効果があるかどうかということは当面の議論とは無関係であるから考慮に入れる必要はない．

［補注４］：χ^2検定やＦ検定では原則として片側確率のみが意味をもつが，検定結果は“Ａ＞Ｂ”ではなくて“有意差あり”という両側検定的な表現となる．ただし１標本の検定でχ^2検定を用いて“Ａ＞Ｂ”と結論される場合や，母分散が既知の値に等しいか否かを検定するような場合を除く．

［補注５］：雌雄の差が見られた場合，“雌雄で有意な差が認められた”と結論すること自体は正しい．また，“当該の行動現象の重要な予測因として雌雄の違いをあげることができる”と述べることも誤りではない．しかし，このことから雌雄の違いが原因であると結論するわけにはいかない．たとえば原因の１つとして，単なるからだの大きさの違いが差をもたらした可能性も考えなければならない．この場合，メスのかわりにメスと同じ大きさのオスを被験体としても同様の差が生じるはずである．いっぽうオスと同じ大きさのメスを被験体とした場合にはもはや“雌雄の差”は見られないであろう．オスのほうがメスよりからだが大きいことは事実であるとしても，“からだが大きい”ことはオスであることの十分条件ではない以上，雌雄の差を原因であると見なすわけにはいかないのである．

［補注６］：個体内比較法のうち単一被験者法についてはバーロー・ハーセン（1993），桑田（1993a，1993b，1994）などを参照されたい．

引 用 文 献

バーロー・ハーセン［著］　高木俊一郎・佐久間徹［監訳］（1993）．*一事例の実験デザイン．ケーススタディの基本と応用．新装版．* 二瓶社．

長谷川芳典（1985）．「血液型と性格」についての非科学的俗説を否定する．*日本教育心理学会第27回総会論文集*，pp.422-423．

長谷川芳典（1987）．血液型と性格 ――公開講座受講生が収集したデータに基づく俗説の再検討――．*長崎大学医療技術短期大学部紀要*，1，77-89．

長谷川芳典 (1988)，2歳児における漢字の読みの学習過程．*長崎大学医療技術短期大学部紀要*，2，139-150．
長谷川芳典 (1989a)．血液型と性格は関係ない！ A link between blood type and personality? Forget it. *THE21*, 1989-2, 86-89．
長谷川芳典 (1989b)．3歳児における漢字熟語の読みと生成．*行動分析学研究*，4，1-18．
長谷川芳典 (1993)．スキナー以後の行動分析学 ——2．心理学の入門段階で生じる行動分析学への誤解．*岡山大学文学部紀要*，19，45-58．
長谷川芳典 (1994)．目分量統計の心理と血液型人間「学」．詫摩武俊・佐藤達哉(編) *現代のエスプリ*，324，121-129．
岩原信九郎 (1964)．*新しい教育・心理統計 ノンパラメトリック法*．日本文化科学社．
岩原信九郎 (1965)．*新訂版 教育と心理のための推計学*．日本文化科学社．
近藤良夫・安藤貞一 (1967)．*統計的方法百問百答*．日科技連．
桑田繁 (1993a)．新しい実験計画法としての単一被験者法の紹介(Ⅰ)——その適用方法と群間比較法との相違——．*全本鍼灸学会雑誌*，43，28-35．
桑田繁 (1993a)．新しい実験計画法としての単一被験者法の紹介(Ⅱ)——データの分析評価法．*全本鍼灸学会雑誌*，43，36-43．
桑田繁 (1994)．新しい実験計画法としての単一被験者法の紹介(Ⅲ)．*全本鍼灸学会雑誌*（印刷中）．
森敏昭・吉田寿夫 (1990)．*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．
岡田泰燊 (1966)．*新しい数学へのアプローチ⑬ 統計*． 共立出版．
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．