

心理学的研究における統計的有意性検定の適用限界

葛 西 俊 治

要 約

心理学における解釈的質的アプローチの妥当性を確認する過程において、統計的数量的アプローチ自体の妥当性及び社会科学への適用可能性が長年にわたって問題視されていることから、統計的検定についてあらかじめ徹底した再吟味を行う必要のあることが明らかとなった。数量的な心理学的調査や実験が専門誌に掲載されるためには、5%や1%の統計的有意水準において帰無仮説が首尾良く棄却されなければならないが、心理学的研究におけるこうした通常の統計的手法については、1970年に書かれた古典的な『有意性検定論争』によって技術的、方法論的、認識論的な批判がすでに行われている。それにも関わらず、1999年の「有意性検定の無意味さ」論文に見られるように、統計的検定は誤って解釈され無意味に乱用されたままであり今日まで改善のきざしもない。

本論文は、1) 有意水準の設定の仕方、2) どういった母集団から標本が抽出されたのか、という二つの基本的な統計的テーマを精査することによって、5%や1%といった「聖なる有意水準」は研究目的に沿って異なるべきであること、及び、母集団の設定については、「人間の斉一性」といった暗黙の想定を回避するために厳密に識別されて定義されるべきであることを見いだした。

また、心理学の歴史とは、個人差や文化的社会的歴史的多様性を無視することによって人間の性質を一般的に捉える方向へと向かう運動であるとともに、知覚心理学や行動科学のように統計的有意性検定を採用することを通じて、いわゆる「科学的学問」へと向かう運動であることが概観された。しかし、a) 抽出された被験者、及び、b) 彼らが属す母集団、に関する属性を明確に定義すべきという技術的要請は、一般意味論が指摘するように容易には達成され得ないことから、厳密科学に向けた一般的結論が有意性検定の結果の解釈によって必ずしも得られるわけではないことが指摘された。続いて、統計学を指向する心理学論文において、統計学的に根本的な問題についての「無記」が見られる現象について議論が行われた。

キーワード：解釈的質的心理学、数量的心理学、統計的有意性検定、有意水準、母集団、標本抽出、人間の斉一性、一般意味論、無記

I. はじめに

心理学の世界は、長年、統計的仮説検定 (statistical hypothesis test) の実施を前提とした数量的アプローチ (metrical approach) を主要な研究方法としてきている。それに対して、近年

のいわゆる「質的アプローチ qualitative approach」は、少数の対象者や特定の一個人を対象とし、面談などの場で得られる言語的資料に基づく詳細な分析によってテーマとする心理的現象や事象に接近する方法として展開されてきた。葛西 (2005)¹⁾はそうした質的アプローチの一つである解釈的アプローチの理論的基盤を示すとともに、「アブダクションに基づくモデル構成法」を新たに提起した。その際、従来の数量的アプローチの適用に関連する問題点として次のような事柄が指摘されていた。すなわち、1) 帰納および演繹という論理的推論が原理的に破綻していること、2) その破綻を取り繕うためミル (J.S.Mill)²⁾は「自然の斉一性」(uniformity of nature)を公理として提起したこと、しかし、3) その公理を人文社会科学の領域での研究に適用するならば「自然の斉一性」の拡大解釈となる「人間の斉一性」といった不適当な仮定を暗黙裏に想定せざるを得ないこと、4) 統計的検定を導入するためには、対象となる現象や事象には反復性や多回性が前提となるため、一回的な現象や少数例の事象は扱えないこと、である。自然科学の領域においては、研究対象とされる物質や物性などが「自然の斉一性」という想定とそれほどの齟齬を来すことなく大域的には妥当すると考えられ、それなりに信頼性の高い研究成果が期待されるのに対して、人文社会科学領域においては、研究対象となる個々人の生まれ育った文化的地理的歴史的経済的等々の背景が異なるため、「人間の斉一性」なる想定は極めて困難といわざるを得ない。そのため、本来、心理学的研究は個々人についての「個性記述的アプローチ idiographic approach」が中心となるべき状況があるにも関わらず、従来の心理学研究はもっぱら数量的アプローチに依拠し統計的検定を前提とする方法を用いてきた。こうした統計的アプローチは、おおむね1940年代から1950年代にかけて成立を見た「行動科学 behavioral science」の実証性および学際性の源となるとともに、心理学全体においても正統的な研究アプローチとしての地位を獲得した。その後、現在に至るまで心理学は統計的検定をほとんど必須の方法として研究の中心に据えてきている。しかし、そうした歴史的経過にも関わらず、心理学的研究において統計的仮説検定あるいは統計的有意性検定 (statistical significance test) を利用するにあたっての技術上、方法論上および認識論上の問題点が様々に提起されてきていることはほとんど顧みられていない。

たとえば、有意性検定の問題について長年研究を続けているミール (P. E. Meehl, 1997)³⁾によれば、アメリカ心理学会 (American Psychological Association) 内には帰無仮説による統計的検定の使用を禁止すべきだという意見とともに、統計的検定の利用は全く問題ないとする者までの両極があること、また、問題視される点も研究者によって様々であることが指摘されている。さらに、そうした状況を改善するには、さしあたり、論文の査読編集者が統計的仮説検定の使用に関するいくつかのルールを採用すれば事足りると述べ、統計的検定に関するいくつかの技術的方針を提示している。だが、統計的検定に対するこうした疑念は、驚くべきことにひとり心理学領域に限られているわけではないことから、単に統計利用に際しての技術的問題としてのみ片付けられそうにない。

たとえば、1999年にジョンソン（D. H. Johnson）⁴⁾によって書かれた“The Insignificance of Statistical Significance Testing（統計的有意性検定の無意味さ）”というタイトルの論文が、ワイルドライフ・マネジメント研究誌の学会賞（The Wildlife Society Award for Outstanding Publication in Wildlife Ecology and Management, 2000）を獲得したという。野生動物や自然生態についての科学研究を行う専門家が自らの研究活動において統計的有意性検定の使用方法には問題があることを指摘し、そのことを明記した論文が学会誌における特筆されるべき論文とされたのである。極めて衝撃的なこの論文は、しかし、有意性検定をめぐる問題は最新のテーマであるどころか、以前から連綿として引き続いてきた問題をあらためて取り上げたに過ぎないことを述べている。統計的検定の利用に関わるこのような議論を遡っていくと、問題意識がそのままタイトルとなった一冊の本“The Significance Test Controversy（有意性検定論争）”⁵⁾に行き着く。1970年に出版されたこの本は、序論の統計的検定に関する説明に引き続き、社会学者19名、心理学者5名、そして関連領域の研究者3名の論文から成り、統計的検定をめぐる当時の論争の詳細をとりまとめたものである。その中には上述のミールによる1967年当時の論文「心理学・物理学における仮説検定：方法論上のパラドックス」⁶⁾も含まれているのであるが、そこでの議論は驚くべきことに近年のミール（P. E. Meehl, 1997）による指摘とそれほど相違はなく、この三十年近い歳月の間、統計的有意性検定に関わる技術的および認識的問題に対してめばしい展開がなかったことを露呈するものだった。すなわち、ジョンソンがすでに指摘していたように、現時点において取りざたされている問題のほとんどが相当以前からすでに論争点となっていたのである。いわく、「（二群の平均値が等しいとするような）帰無仮説（null hypothesis）は、元々採択される可能性はきわめて低く、必然的に棄却されるべきものである」、「したがって点推定的な帰無仮説ではなく区間推定を用いるべきである」、「帰無仮説が棄却できないのは標本数が少なすぎて検定力（test-power）が不足しているためである」、「 χ^2 検定やt検定では、標本数が非常に多い場合は、調査や実験をするまでもなく必ず有意となる。そのため、どんなに小さな平均値の差でも有意となるから、統計的に有意となったからといっても実質的な意味が薄い」、「有意水準に達しない研究は心理学専門誌に投稿されず投稿されても採用されない」、「そのため、慣行として用いられている有意水準の5%や1%という確率値がほとんど神格化されている」などといった技術的指摘とともに、「帰無仮説の棄却によって対立仮説の証明はできない」「統計的検定は集合的なデータについての判断であって、一般化に向けた判断とは異なる」などの認識論的な指摘も行われていた。

この長い歳月の間にたとえば心理学者たちは、数量的アプローチの根幹ともいえる統計的検定について、一体どの程度にまで理解を深めてきたのだろうか、あるいはどれほどの無理解に陥っていたのだろうか。そして、その理由は一体どこにあるのだろうか。心理学者を含む社会科学領域の研究者に対して統計学的素養の不十分さを指摘する向きもあるにしても、ジョンソンの論文が指し示すように自然科学領域の研究者も同様の問題を乗り越えられていないと

いう事実もある以上、問題は統計学についての素養や理解力といった側面にあるというよりも、むしろ仮説検定についての研究者側の認識ないし心理的な要因、あるいは研究者集団や学会における研究論文の投稿や採択に関わる社会的な要因が関係していると考えるべきであろう。

ところで、統計的検定に対するこのような無理解の歴史とは反対に、データを解析するための統計学的分析手法はこの歳月の間に大きな展開を遂げてきており、その成果は心理学の世界にも取り入れられ広く用いられてきている。『統計的有意性論争』が出版された1970年当時はまだそれほど一般的ではなかったがその後の主要な多変量解析プログラムに盛り込まれた手法としては、たとえば、コンジョイント分析、多次元尺度構成法、共分散分析、対数線型モデル、共分散構造分析等が挙げられる。特に近年、重回帰分析と因子分析との結合とでもいえる共分散構造分析（あるいは構造方程式モデル SEM : Structural Equation Model）の発展の度合いは目覚ましく、今後、人文社会科学領域における中心的な分析法ないし研究法の地位を獲得するとも言われている。こうした分析手法の発展とも相まってか、あるいは高度情報技術に基づく医療機器の発展の故なのか、たとえば医学領域においては明確な「データ（証拠）」に基づいて医療的対処を行うべきであるとする EBM（Evidence Based Medicine）という立場が大きな勢いをもつに至っている。さらに、複数の研究における統計的結果を総合して現象を大域的に把握するための手法「メタ・アナリシス meta-analysis」への関心も高まるなど、情報処理技術としての統計学はすでに方法論上の中核となって久しいといえる。

こうした時代的な勢いの中にあっては、統計的有意性検定に関する本論文における議論は一見些末なものに見えるかもしれない。しかし、アメリカ心理学会において統計的検定を禁止すべきだという立場すらあるということや、ワイルドライフに関する科学的研究論文が統計上の問題を依然として抱えたまま論文誌に掲載され続けている実態が指摘されただけではなく、そうした問題群が30年以上にわたってほとんど解決を見ることなく引き続いてきているということは、統計学的アプローチの、そしてそれを基本的方法としてきたおおかたの心理学的研究ならびに社会科学全体にとっても極めて憂慮すべき事態と言わざるを得ない。

さらに、心理学領域において近年発展を続けている、いわゆる「質的アプローチ qualitative approach」あるいは「解釈的アプローチ interpretative approach」という方法論を心理学領域において明確に定位させていく過程では、従来の数量的アプローチとの方法論上の対比は必須のこととなる。その際、主に言語的な資料を解釈的に用いていく質的アプローチあるいは解釈的アプローチに対して、従来の数量的アプローチに基づいた立場からは、主観的であるとか厳密さを欠くなどといった批判が加えられることがある。しかし、統計的検定をめぐる論争が依然として存在している以上、質的アプローチへ対しての批判を加えるに足るだけの適格性や妥当性を現行の数量的アプローチがどの程度備えているのかという新たな論点も浮かび上がってきたといえる。

本論文は、そうした状況に関する問題点を明らかにし詳細な吟味を行うことによって、心理

学を含む社会科学の領域においては、統計的検定によるアプローチは一定の制約の下においてのみ用いられるべき限定的なアプローチであることを明らかにする。さらに、統計的検定による結果の解釈が、そこに至る数量的アプローチによる計算上の厳密さとは裏腹に、後に示すように何らかの「比喩的解釈」に基づいてなされていることを指摘する。

以下では、まず第一に、心理学における統計的検定に基づくアプローチの問題点について概略を述べ、そうしたアプローチの適用に際して想定されている暗黙の前提を明らかにしていく。初めに、統計的検定や確率に関わる技術的な問題点にふれ、次いで、統計的検定の利用に関わるいくつかの根本的な問題点について吟味していくことにする。

Ⅱ. 「自然の斉一性」原理から統計的検定へ

帰納の原理的問題

いくつかの個別的知識から一般的法則を導き出すことを帰納（induction）という。たとえば、「手を放すと物が落ちる」という命題は、それを何度も試行し確かに落ちることが確認されることによって、一般法則として真であるとされる。また、そのようにして得られた一般的な命題に基づいて、「手を放すと、持っている物は落ちるだろう」というように起こるべき現象の予測をたてることを演繹（deduction）という。科学は、反復実験を行う帰納法によってこうした真なる命題を打ち立てるといわれるが、よく考えてみるならば、そうしたことは原理的には不可能であることに気がつく。すなわち、何万回試行しても次の回には「落ちなかった」という事実が起きるかもしれないという可能性を排除できず、そうした可能性を排除するためには無限に試行を続けなくてはならない。しかし、無限に試行を続けるために、いつになっても一般化に達することができず、命題はいつまでも蓋然的命題に留まり続ける。そのため、帰納によって普遍的命題に至ることは原理的に不可能と営みとなる。したがって、帰納法に頼ることによって一般法則に至るというのは単なる希望的観測や信仰に過ぎないと言わざるを得ない。

こうした難題に遭遇したミルは、「自然の斉一性 uniformity of nature」という概念を「公理」として持ち出すことで乗り越えようとした。つまり、自然の性質は斉一（均一、一定不変）であるので、「落ちる」といった現象が何度も起きるのなら、それは自然の性質としてそうなのだという「公理」において断定するものだった。こうした理解はすでに葛西（2005）において示されている内容である。さて、問題はそうした「自然の斉一性」といった対処とは直接的な関連のない新たな学問領域である統計学において、ミルによる公理と対比されるべき発想が生み出され、それが数量的アプローチの根幹に据えられていった点にある。

統計的検定の基本的な内容を眺めてみるならば、厳密に見える統計的アプローチは、その設定と計算方法については数学的に厳密であるけれども、有意水準の設定自体は以下に詳述するように人間的スケールでの常識的判断に基づいているのに過ぎない。そのため、複数の個別事

例の積み重ねによって一般的命題へ至ろうとする帰納という方法について、その原理的問題点が統計学の登場によって解消された訳ではないことは明確にしておかなければならない。したがって、数量的アプローチは、ミルが「自然の斉一性」という公理を導入せざるを得なかったという歴史的事実からも明らかなように、第一点として「帰納の原理的妥当性の欠如」、そして、第二点として新たに「統計的有意水準を設定するための原理的基準の欠如」という二つの限界を抱えていることを以下において議論する。

なお、帰納における原理的困難さは、ポパー (K.R.Popper)⁷⁾ による「反証主義 falsificationism」を生むきっかけともなり、統計学においてもそれと同等の論理にもとづいて仮説検定の方法が設定されている。すなわち、あらかじめ否定されるべき「帰無仮説」を提示しておいて、それが棄却されることによって結果的に「対立仮説」(本来、証明したい仮説)を採用するという推論形式である。しかし、この方式にしても積極的に「対立仮説」を証明することに成功したわけではないことに注意する必要がある。つまり、主張したい仮説が大手を振って証明されたのではなく、主張に対抗する(帰無)仮説が否定されたために、帰無仮説と対立する仮説の蓋然性が高まったというのが統計的仮説検定の骨子だからである。したがって、そうした二重否定的な了解にどれだけの妥当性があるのかという問題は依然として残されたままといえる。

さて、以下では、まず数量的アプローチがその土台とする「統計的仮説検定」に関わる二つの技術的問題を取り上げる。一つは、「無視できるほどに小さな確率の値」の設定による有意水準 (level of significance) とそれに関連する問題である。二つ目は、仮説検定を行うために必要な観測データ数ないし被験者数・回答者数に関わる「標本数 sample-size」の問題と、それによって推計される「母集団 population」に関する問題である。特に後者においては、統計学上のそうした技術的問題点に端を発して、認識論的な難題に直面せざるを得ないことへと言及する。

1. 統計的有意水準に関連する問題

帰無仮説とは

統計学は「自然の斉一性」といったような公理を立てるのではなく、ある現象が確率的に極めて小さなものである場合「それは起きないものとする」という立場をとることによって、何らかの命題や事態の(いわゆる)「一般性」を主張するものとなった。すなわち、「たまたま起きたに過ぎない、可能性の小さな事象」を理論的に排除することによって、主張したい命題や事態の妥当性を確保するだけでなく、無限に試行を繰り返さざるを得ない帰納法の問題点を易々と乗り越えたかのようであった。そうした魔法のような力が、実は「p値」と呼ばれる確率値、すなわち $p < 0.05$ 、「確率5%以下」ないし $p < 0.01$ 、「確率1%以下」といった、一見ささやかな確率数値指定によって生み出されている。この数値を基準として研究上の仮説の採択ないし棄却という決定が行われることから明らかなように、統計的検定の実質はまずはこ

の数値の設定に掛かっている。判断の基準として通常用いられている5%あるいは1%といった確率値のことを有意水準（level of significance）という。また、帰無仮説の下で当該のデータが（所定の分布のもとで）得られる確率値のことを「P値 p-value」と呼ぶ。データから計算されたP値が、設定していた5%ないし1%の有意水準よりも低いとき、それを「 $p < 0.05$ 」ないし「 $p < 0.01$ 」と表記し、帰無仮説が棄却されることを示す。このとき、「仮説 H_0 は、5%水準（ないし1%水準）で棄却された」というように表現して、ある仮説 H_0 が発生する確率は5%以下、すなわち、「それはかなり小さな確率でしか起きないことだから、たまたまそういう小さな確率の事象が起きたというように考えるのは合理的ではない…」という判断を行う。そのようにして仮説 H_0 を拒否・棄却（nullification, rejection）することによって、結果的に「それと対立する仮説 H_1 を採択する」といった論理展開をみる。このように、当初から棄却することを狙いとして提出される仮説 H_0 を帰無仮説と呼び、それを却下することによって、帰無仮説と論理的に排反関係にある仮説、すなわち、対立仮説 H_1 の妥当性を確保しようとする。なお、わざわざ否定されるための帰無仮説を設定する理由は論理学の次のような基本に基づく。すなわち、「 $A \supset B$ 」（AならばB）で今「B」であったとき、「 $B \supset A$ 」（BなのでA）とはだちに主張することができず、主張すれば「後件肯定の誤り fallacy of affirming the consequence」となってしまうことによる。つまり「雨が降れば→道路が濡れる」とき、「道路が濡れている」から「雨が降っている」とはならないということである。したがって、「 $A \supset B$ 」を証明するためには、真となる論理展開であるその対偶（contraposition）、「 $(\neg B) \supset (\neg A)$ 」、すなわち、「BでないならばAではない」という論理を用いることになる。つまり「道路が濡れていない」のだから「雨は降っていない」と正しく主張できるとされる。（なお、このアリストテレス的二値論理は一般意味論によってすでに疑問視されているが、ここでは一般的な想定を示す）。

したがって、統計的有意性検定とは、まず、本来証明したい「（対立）仮説」を設定すると同時に、それとは反対の内容となる「帰無仮説」を設定する。後者は否定されるために出された技術的な仮説であり、すでに示したように対偶の考え方をを用いて対立仮説を採択するための道具立てとなる。また、この帰無仮説によって一定の確率分布が想定されるため、当該の研究によって得られたデータが、帰無仮説のもとでどの程度の確率（いわゆるP値）で起きるかを計算することが可能となる。次に、帰無仮説を棄却してもよいと考えられる確率値5%や1%を有意水準として設定する。そして、データ分析の結果、帰無仮説の下でそうしたデータが得られる可能性が5%や1%よりも小さな確率だったときには帰無仮説を棄却し、本来証明したかった対立仮説の妥当性を確保することになる。なお、これによって「対立仮説が真である」ことが証明されたというよりは、厳密に言えば、有意水準の確率値に相当する誤りの危険性を前提とした上で「対立仮説の蓋然性が高まった」ということである。そのため、こうした論理展開によって棄却される帰無仮説と採択される対立仮説は、構造的に二種類の確率的な誤り、

「第一種の誤り・第二種の誤り」を含むものである。このように、帰無仮説に基づく統計的検定は、実質的には有意水準と呼ばれる確率値をめぐる次のように進められることになる。

有意水準という確率値

小さな坂で鉄球を放すと鉄球は坂から平地まで転がって行き、例えば1メートルほど先まで転がっていったから止まるという状況でこのことを実験的に調べてみることにする。何度か行くと鉄球の止まる場所はほとんど同一であることが分かるが、それと同時にその位置よりもわずかに手前に止まったりほんの少し先まで行ったりするという誤差も認められる。鉄球表面のデコボコや床の状態、風の有無や磁気など多くの攪乱要因による影響があるためである。しかし、ひどく手前に止まったりズーッと遠くの方まで転がっていくことは極めて稀である。そのように平均的に止まる位置を中心に、その位置からどの程度ずれて止まるかという誤差とそのような事例が起きる確率が「正規分布にしたがう」ということが統計学的知見として得られている。「鉄球がズーッと転がっていき平均値を何十センチも越えた所で止まる」というような通常は考えられない事態について、統計学的判断では、そうした事態が出現する確率（P値）がひどく低いから「起こらない」と判断する。そして、そのときの判断基準とされる確率値は通常は「5%以下あるいは1%以下」と設定されることがほとんどである。なお、歪みのないサイコロを転がしたとき、「一の目」が二回続けて出る確率は $1/36=0.027$ で約2.7%、「一の目」が三回続けて出る確率は $1/216=0.0046$ で約0.4%と計算されるから、有意水準とされる確率値が「5%以下あるいは1%以下」というのは極めて小さな数値とは言えず、現実的に体験できる程度の設定と言えるだろう。なお、統計学では「自然の斉一性」を公理として採用したのではなく、「想定した数値よりもひどく小さな確率だから、その現象は起こらない」という判断を、統計学上の公理としている。しかし、そうした判断に関わる根本的な問題は、「小さな確率」とされる5%なり1%なりという数値が実は恣意的に設定された値であってそれ自体に原理的な必然性がない、という点である。

なお、5%という確率値は、農業と生物学に関する実験の場面において、統計学者 R.A.Fisher の著書 “The Design of Experiments” (1935) の中で初めて用いられたとされる。（“The Significance Test Controversy” p.156, 1970参照）

さて、鉄球を坂で放して転がす実験を多数回行ってみた結果、鉄球が止まるまでの距離の平均値が100cm、標準偏差が1cm だったとすると、正規分布の知識を用いることによって、99%の確率で $100\text{cm} \pm 2.57\text{cm}$ 、すなわち、97.43cm から102.57cm の間で止まること、また、95%の確率で $100\text{cm} \pm 1.96\text{cm}$ 、すなわち98.04cm から101.96cm の間に止まることが計算される。しかし、実際にはそれよりも小さな確率事象が起きる可能性があるわけで、どの程度にまで「慎重」であるか「剛胆」であるかといったように情緒的に理由を述べる代わりに、慣習的に1%

なり5%なりの数値を「有意水準」として技術的に設定したのに過ぎない。

ところで、ゲーム理論の考えを方を最初に提示したと言われるフランスの数学者ボレル（E.Borel）によると、宇宙的水準では 10^{-500} の確率は無視できること、地球的スケールでは 10^{-15} 以下の確率は無視できること、そして人間的スケールでは $10^{-6}=0.000001$ 、すなわち百万分の一以下の確率は無視できるという⁸⁾。百万分の一、すなわち0.0001%などという極めて小さな確率値は、シックスナイン（99.9999%）と呼ばれる精度にまで故障などの事例を許さないという数値でもあるから、限りなく起こりえない小さな確率値に見える。実際、工業製品などは自らの精度の高さを謳う際に示す数値がこの99.9999%というものとなっている。そうした小さな数値ではあるが、意外にも思った以上に実際的な数値でもある。たとえば、アメリカのスペースシャトルのように複雑な機体では全体を構成する部品点数が極めて多いため、部品点数がたとえば約69万点のレベルを超えるならば、それぞれの部品の故障率が百万分の一だったとしても、部品のどれかが故障する確率が50%を超えると計算される程度に現実的なのである。事実、個々の部品やユニットのトラブルの確率が極めて低いものであっても、巨大なシステムがしばしばトラブルに見舞われるのは、一つにはこうした部品点数の多さという単純な事実に依ることが知られている。そうした状況においてエラーを極力抑えようとするとき、いわゆるシックスナインという精度はおおむね技術的に実現可能な限界値なのであって、それさえ実現できればエラーを押さえ込めるという訳ではない。そうした点では、一見して極小に見える百万分の一（0.0001%）という確率値といえども、巨大な機械システム装置などにおいては人類が経験し得る程度に「人間的スケール」での確率と考えられる。ちなみに、建設当時、数万年に一回程度しか起きないとされていた原子力発電所の大規模事故は、現実には、1979年のアメリカ・スリーマイル島での炉心崩壊事故に続いて1986年のウクライナ・チェルノブイリ原子力発電所での炉心爆発事故と、地球的スケールならぬ人間的スケールにおいてすらいわば「立て続けに」起きるなど現実のものとなっている。（巨大システム自体の問題性とそこで働く人間のヒューマンエラーとの組み合わせが事故の要因とされている⁹⁾。）

確率値をこのように捉えることで百万分の一（0.0001%）程度の極小確率であってさえも「人間的スケール」内にあるとするならば、p値として用いられてきている5%なり1%といった数値は実に驚くべきほど高い確率と言わざるを得ない。それにも関わらず、そうした確率値を仮説の棄却に際して用いるというのはどういう理由によるのだろうか。

たとえば、1970—80年代頃に盛んに行われていた主観的確率の研究に関連して、葛西（1989）の「低確率事象の認知研究」¹⁰⁾によれば、人間の確率判断（主観的確率）は極めて不正確であり、人は低確率事象を正当に評価する言語的認知的システムを欠いているという。したがって、統計的検定における「有意水準」とは、統計学の学問上の厳密性とは関係なく、人間の低確率事象についての認識力の低さゆえに、「1%ないし5%」といったような「極めて寛大な」確率値を有意水準に設定したと推測されるところではある。

なお、ここで5%ならびに1%という数値に対して「極めて寛大」と述べた意味を、機械部品の例で示しておく。いま5%という確率値をたとえば機械部品の故障率と考えてみる。すると、故障率5%の部品が13.5個以上から構成される製品があったならば、その製品の部品のどれかが故障を起こす確率は単純計算で50%以上となる。これは驚くほどに不安定で使用に耐えない製品といえよう。同様に、部品の故障率が1%である場合は、そうした部品68.9個以上から構成される製品は、その部品のどれかが故障する確率が50%を超えることが計算される。この状態は5%の場合よりはまだまだしであるが、精密な機械システムが要求する故障率0.0001%による場合と比較すると明らかに不安定な製品と言わざるを得ない。

これを心理学研究の場面に適用してみよう。有意水準5%で帰無仮説が棄却されその対立仮説が採択されるとして、そのようにして分析結果を得た心理学論文が専門機関誌の一冊に13-14編ほども掲載されていたとするならば、その中のどれかの論文が誤りである確率が何と50%を超えるという状況を意味する。すなわち、それぞれの論文の正しい確率が95%であるから、その全ての論文が正しい確率は0.95を13乗ほどすると0.5以下となるということである。また、ある心理学専門誌の論文がすべて有意水準1%に基づいて仮説を採択したとするならば、そうした論文が年間に69編以上も掲載されたとき、その中の論文のどれかが誤りである確率は50%を超えると計算される。あるいは、年間に100編もの心理学論文（有意水準1%によるもの）が掲載されたとして、掲載論文がすべて正しい確率は約36.6%程度しかないという驚く程に小さな数字となる。このように、「20回に一回…」である5%や「100回に一回…」である1%という有意水準は実は極めて大雑把な数値であって、それによって帰無仮説を棄却するということは、発表論文全体として眺めてみるならば、「その程度の精度」によって研究を行っているということであり、そうした論文を掲載する研究誌は全体としてみるならば、厳密さをそれほど主張できない状況にあることを明確に認識する必要があるだろう。

さて、百万分の一、すなわち0.0001%というところまで極端に「臆病」にならずに、仮に「万に一つ（一万分の一）」もミスがない程度に「慎重な場合」、0.01%を有意水準として用いて99.99%の場合までカバーするものとなる。これは例の鉄球転がし実験においては96.11cm～103.89cmという範囲に収まる、といった数値である。あるいは「五分五分」などのように「極めて剛胆な場合」には、50%という数値を有意水準とすることで現象の50%をカバーすれば良いのならば、同じく鉄球が転がって止まる位置は、平均から±0.33cm、すなわち99.33cmから100.67cmまでとなる。このように様々な「有意水準」が想定可能であるため、検定の際の慣習の設定値とされている1%ないし5%という数値の甘さに疑問をもち、当該研究者集団が「きわめて慎重」な態度を取ることに決めたでしょう。そして、有意水準を仮に「万に一つ」のレベルである0.01%に設定したのならば、今日まで積み重ねられてきた数量的アプローチに基づく心理学研究のほとんどはそこまで「慎重な」基準には届かず無効な研究の山と成り果て、そうした心理学分野はほとんど壊滅状態となるだろう。つまり、これまでの数量的な心理学研

究が一見「厳密な研究」として位置づけられてきた一つの理由は、特に根拠もなく用いられている5%なり1%といった「かなり剛胆」な有意水準の設定によるものであり、そうした限定的な妥当性の範囲内での主張に過ぎないものであることが分かる。有意水準を0.01%程度にまで低く設定するだけで、当該研究領域の知見がほぼ壊滅し得るということから明らかなことは、有意水準の数値をどこに設定するかという問題は当該学問領域のあり方を決める極めて重要なパラメータだということである。したがって、それを5%や1%といった慣習的な数値に設定しておくことが当該学問領域において確かに妥当なのかを根本的に議論しなければならないだろう。

ところで、市販の関数電卓においても確率分布の数値が容易に計算され、統計ソフトは確率値（P値）を当たり前に表示する現代とは異なり、統計学が発展を遂げる一世紀ほど前は、確率値をあらかじめ手計算してその結果を確率数値表（正規分布、t分布、 χ^2 分布、F分布…）として提示する以外に実際的な方法がなかった。そのため、ありとあらゆる条件に備えて無数の確率表を用意するわけにはいかず、たまたま1%や5%などといったような切りのよい数値の確率表が作成されて用いられたという背景事情があったと推測される。したがって、研究の結果、 $p < 0.049$ というP値が得られれば「有意水準の5%以下」ということで仮説の妥当性が得られたと小躍りし、結果が $p < 0.051$ というP値ならば「帰無仮説が棄却できない」として落胆するといった程度にまで「有意水準の5%」やらあるいは「1%」といった数値自体に絶対的な意味があるとはいえない。また、「有意水準を5%とするか1%とするかを事前に想定」しておいて、研究結果においてたまたま $p < 0.023$ というP値が得られたとして、有意水準を5%に設定する心づもりをしていたので安堵するとか、あらかじめ1%にしようと心づもりをしていたので落胆するというのも統計学的には非本質的なことといえよう。（この例は「心づもり」という主観的なプロセスについてのものである。後に示すように、第一種の誤り、第二種の誤り、サンプルサイズの三つの項目をどのように定めて研究を設計するかという意味での記述ではない。）

さて、ここではまず、統計的検定に用いられる有意水準が、「研究の厳密さという意味では5%や1%どころか0.01%なり0.0001%なりの極めて厳格な数値であるべき」という方向性が示されたが、ここでの議論はあくまでも5%や1%という確率値がどの程度の曖昧さを伴っている数値であるかを示すためのものであって、有意水準をそこまで小さな確率値にすべきであると主張するものではない。というのは、実はそれとは全く逆に「有意水準は5%や1%などの低い数値ではなく、10%や30%などの大きな数値にするべきである」という正反対の設定を行うべき事情が存在しているからである。なお、先取りして述べるならば、こうした議論の要点の一つは、有意水準という数値の基本的な不定性を明確に示すことにあり、慣習的に用いられている固定的な有意水準、たとえば「聖なる5%」への神話をさしあたり打破することにある。

第一種の誤り・第二種の誤り

『有意性検定論争』の「聖なる.05」という一章において、社会科学では慣習的に5%, 1%, 0.1%が用いられているが、その中でも特に“0.05”が最も「聖なる sacred」数値となっていると揶揄されている⁵⁾ (pp. 155–160)。この数値は有意水準と呼ばれる同時に、いわゆる「第一種の誤り」の度合いを設定する数値ともなっていて、統計的検定では必須の項目である。ちなみに、統計的仮説検定には「第一種の誤り Type I error」と「第二種の誤り Type II error」といった「二つの誤り」が構造的に存在している。「第一種の誤り」とは、「帰無仮説が正しいのに誤って棄却してしまう」ことであり、この確率の大きさを α で表す。このため α 過誤と呼ばれることがある。続いて、「第二種の誤り」とは、「対立仮説が正しいのにそれを採用しない誤り」であり、この確率の大きさを β で表す。このため β 過誤と呼ばれることがある。なお、 $(1 - \beta)$ によって表される確率値は「対立仮説が正しいときにそれを正しく採用する」確率値となるため、「検定力 test-power」と呼ばれる。さて、これまで述べたきた有意水準とは、この第一種の誤りの大きさを示す α と一致する。というのは、帰無仮説が棄却される確率が5%なり1%なりの数値であるとき、データから算出されたP値がこれよりも小さい場合には帰無仮説が棄却されるのであるから、その確率値の分だけ誤って帰無仮説を否定しまう可能性（第一種の誤り）があるからである。

すでに述べたように、有意水準すなわち α を0.1%や0.001%などの小さな数値をすれば、帰無仮説を誤って棄却する可能性はその数値にまで低く抑えることとなる。先に示してきたP値に関する議論はそうした観点に立って進められてきたが、もちろん、第一種の誤りの可能性である α を極めて小さく設定するのが最善というわけにはいかない。なぜなら、いつまでたっても帰無仮説が棄却されないでいるという可能性が高まってしまうからである。これは事情によってはかなり悲惨な結果を引き起こしかねない事態となる。たとえば、ある症状を改善する新薬が開発されたとして、その効果の判定のために帰無仮説を棄却する有意水準を仮に極めて小さな数値、たとえば0.001%としたとしよう。研究上の仮説としては、たとえば、偽薬 (placebo) を投与された対照群と新薬を投与された実験群とで、「症状の改善効果が等しい」とするのが帰無仮説となる。もちろん、対立仮説である「新薬による症状改善効果が認められる」ことを主張するのが目的である。さて、実際には新薬が効いていてそれなりに症状が改善されていたとしても、誤って帰無仮説を棄却する確率が0.001%といった小さな確率になるまで、つまり、「偽薬による効果と比較して、新薬の効果が極めて明白」となるまで「帰無仮説を棄却しない」のであるから、新薬によるよほどの効果が確認されるまでは「偽薬も新薬も効果が等しい」という帰無仮説は棄却されないままとなる。したがって、実際には症状改善の効果がそれなりに確認されたとしても、統計的検定上は「新薬によって症状改善効果があった」とする対立仮説は採用されず、新薬は承認されないことになる。これは「第一種の誤り」の確率 α を可能な限り低くしようとしたために、「第二種の誤り」の確率 β が極めて大きいものとなる事態である。

さて、これとは逆に、実際はほんの少しでも症状改善が見られるのならば新薬として承認する方が良いでしょうから（重大な副作用がないとして）、帰無仮説を棄却する有意水準を0.001%などの小さな数値に設定するのではなく、その反対に例えば10%とか30%とかの大きな数値に設定しておくとしよう。すると、新薬の投与によって少しでも効果が見られたならば容易に帰無仮説を棄却することが可能となり、それによって対立仮説「新薬には効果がある」が主張できる。こうした設定によれば、「第二種の誤り」の確率 β は小さい数値に抑えられる反面、「第一種の誤り」の確率 α が相当に大きくなる事態である。

こうした例から推測されるように、「帰無仮説を誤って棄却する確率」である有意水準をたとえば10%なり30%なりに高めに設定することにはそれなりの現実的な意味がある以上、5%なり1%なりに固定してしまうべき理由は希薄となる。さて、第一種の過誤（ α 過誤）と第二種の過誤（ β 過誤）、そしてサンプルサイズ（標本数）という三者の関係をどのように設定すべきかという統計技術的テーマについて、永田（2003）¹¹⁾は次のように簡明に述べている。

検定における2種類の誤りに関する基本的事項（引用）

- (1) 帰無仮説は限定的なので、第一種の誤りの確率（有意水準） α は一つに決まる。
- (2) 対立仮説は複合的なので（パラメータがいろいろな値をとりうるので）、
第二種の誤りの確率 β はパラメータの値が異なると変化する。
- (3) α は小さく設定できたとしても、 β は非常に大きな値になりうる。
- (4) α を大きくすると、 β は小さくなる。
- (5) サンプルサイズが大きくなると β は小さくなる。

これと同様の内容は山内（1998）¹²⁾によって書かれた、心理学および教育学研究者向けの統計的アプローチの解説書にも示されているが、それほど馴染みのあるものとはなっていない。さて、第一種の過誤と第二種の過誤については、その両者は逆向きの関係にあるために一方の誤りを避けようとすると他方の誤りの可能性が高まることになる。そのため、（標本数を増やすという方法をとらなければ）二つの誤りの可能性を同時に低減させることはできず、いずれにしてもどちらの誤りの可能性を低くすることを優先するかという判断が要求される事態となる。こうした第一種の過誤と第二種の過誤の兼ね合い、そしてサンプル数を確保するための時間的経済的といった現実的な制約の中で、研究者は統計的検定に至る基本内容をどのように設計するのかという研究上の判断が求められているといえる。また、第一種の過誤と第二種の過誤の確率値をどのように設定するかは、信号検出理論（signal detection theory）¹³⁾で示される考え方と同様に、それぞれの過誤の確率とそれに伴う効用（utility）ないしマイナスの効用（disutility）とを積和して算出される期待値効用を最大化することに基づいて判断することが考えられる。この場合も、それぞれの数値が自動的に決定される訳ではなく、当該研究者によ

る「判断」によって、用いるべき数値が設定される事態なのである。

ここにおいて、統計的検定には、統計学の知見そのものと、その知見を特定の研究領域で用いる際の当該研究者の介在という二つの要素があることが明らかとなる。つまり、統計的有意性検定の過程は単なる自動的統計計算ではなく、山内(1998)も指摘するように、計算に用いられるべき「過誤」の度合いを設定するという、当該研究者による主体的な関与を必要とするのである。したがって、仮に1%なり5%なりの数値を有意水準として設定したのならば、当該研究者は、その研究においてそうした数値を用いる根拠を明確に示さねばならないだろう。しかし、実態は心理学的研究の多くが慣例的に $p < 0.05$ や $p < 0.01$ というp値を論文に載せるだけであってそうした記述の前提となる有意水準の設定の理由を示すこともなく、また心理学領域の論文審査の査読者や編集委員もそうした指摘を系統的に行ってきたという形跡もない。このように、有意水準の確率数値設定に関する理由を問わない研究誌の編集実態によって見過ごされてきた事柄が、統計的有意性検定についての技術的問題の一定の部分を含めていられる。したがって、この30余年、見過ごされてきた有意性検定をめぐる問題は、さしあたり次の要請によってある程度の改善を期待することができるであろう。

要請1 統計的有意性検定を用いる心理学的研究に対して、本論文は次の三点に関する記述を明示すべきであることを主張する。

すなわち、1) 有意水準とされる数値が5%や1%などではなく当該研究テーマに即して吟味されているか、2) 対立仮説が誤って採択されないままとなる β 過誤についての配慮は適切か、3) 適切なサンプルサイズとなっているか、の三項目である。

もしもそうした記述がなされなければ、検定結果の位置付けに疑義があることから、当該研究論文自体の妥当性が問題となるだろう。なお、繰り返して指摘しておくが、統計的有意性検定をめぐるこうした観点はすでに統計学そのものの問題ではなく、例えば心理学という学問領域において統計的検定を用いる際に、研究領域やテーマに即して上記の三点はそれぞれどのようなものであるべきかといった吟味と周知に関わる、心理学的研究の方法論上の問題なのである。

さて、有意性検定に関わるこうした技術的問題についての周知徹底が図られることによって状況の改善が期待されるけれども、極めて不思議なことは、この程度の技術的改善策が心理学の専門家集団によって長年受け入れられなかったのは結局どういう理由によるのかが不明な点である。つまり、ミール(1967)から三十年近く経過してなお同様の批判がミール(1997)の論文に見られることの原因は何か、ということである。すでに述べたように、統計学的手法を研究法の中心に据えている心理学は高度な統計学的手法を次々に取り入れているのであるから、上に示した三項目程度の技術的指摘はそれほど問題でもないと考えられるためである。

この点について熟考してみるならば、確かに、有意性検定に関わるこうした技術的側面そのものが問題だったのではないという、もう一つの可能性に行き当たる。すなわち、こうした技術側面が長年の間、特に問題として取り上げられないでいるという状況を作り出している何らかの要因や事情が介在する、ということである。さらに、ここで指摘した程度の統計学上の内容を何故か見過ごしたり見落としをさせてきた理由や原因を明確にしなければ、結局は1970年の『有意性検定論争』当時の議論のまま、何らの進展もないまま次の数十年が経過することにもなりかねない。

したがって、有意性検定に関わる問題は、第一には、長年指摘されてきたにも関わらず今日まで見過ごされてきた統計学上の技術的側面にあること、そして、第二には、そうした見落としを長年にわたり系統的に招いてきた理由や原因にその本質が潜んでいること、としてあらためて捉え直さねばならないのである。それでは、後者の「本質的な理由や原因」とはいったい何であろうか。1970年当時からの議論には、有意性検定をめぐる「認識論的」側面からの議論も散見され、「ベイズ統計学」「反証主義」「（後件肯定の誤りなどの）論理学」などの論点はすでに示されてきている。また、ミール（1997）による論文は極めて端的にタイトルそのものが“The Problem Is Epistemology, Not Statistics:...”，すなわち「問題は認識論的であって統計学的ではない」ことを謳っている。しかし、その内容は、仮説検定一般に関わる論理構造についての議論が中心を占め確かに認識論的ではあるが、それによって要請1にある技術的問題についての無視が解消されるといった性質のものではない。また、ジョンソン（1999）は、主に有意性検定の技術的問題点という事実を提示し、論文査読者と編集者に対して科学の裁定者（arbiters of scientific practice）足るべきことを述べる程度である。

そうした現状の中、本論文では、次のような事柄についての考究を通じて、確かに「認識論的な問題」が存在し、それによって有意性検定に関わる技術的問題が潜在化されたと捉えるに至った。まず第一に、統計的検定が、歴史的には「事例を列挙することによって一般化に至る」推論形式である帰納法の代替的方法として採用されてきたこと、第二に、心理学の歴史において、いわゆる自然科学的方法論への希求が統計的検定をいわば神格化の位置まで高め、個性記述的な学問的方法論を貶める経過をたどったこと、第三に、そうした心理学の方向性が、以下に示すように、研究に関連する「観測数ないしサンプルサイズ」の設定において根本的な問題をはらむことになったこと、以上三点の複合的な働きによって、有意性検定の技術的問題が潜在化されたことを指摘するものである。

さて、以下に示す「観測数」に関する議論は、さしあたりは統計学の技術論的な「サンプルサイズ」および「母集団」をめぐるものであるが、それらの概念をどのように捉えるかという「認識論的」な側面に本質的な問題が潜んでいることを明らかにしていく。

2. 観測数と母集団に関わる問題

複数性という要件と統計学的抽象

数量的アプローチには、統計的検定によって仮説の妥当性を求めるという基本的構成のために、暗黙のうちに自明視されている方法上の原則、すなわち、複数回試行する実験や複数人からなる集団や階層の存在など、常に複数性によってのみ担われるという原則がある。つまり、反復が不可能な一回的な状況では基本的に統計的手法は不可能であり、また少数の事例や少人数の場合には統計的方法によって成果を期待することはできないのである。ところで、個々人の生きる世界は本質的に一回的であるから、被験者数の多少に関わらずそのままでは数量的アプローチの要件である複数性という概念には至らない。そこで、個々人の一回的な状況を抽象化し個別性を排除するという、研究者による概念操作が必須となる。すなわち、個々人はそれぞれに独特な生まれ育ちを経てきた特異性に満ちた存在であるために、そのうちの特定の属性（性別、年齢、職業、居住地などや研究対象となっている調査項目や測定項目）にのみ着目し、それ以外の個別性をすべて無視ないし排除し研究対象から除外するという概念的操作を行う。つまり、ここでは研究者側の関心によって注目されている属性だけが意味ある属性とされ、個別性に関わるそれ以外の次元や属性はすべて、テーマとする事柄には直接の関係を持たないが統計的に影響を及ぼしうる「誤差」ないし「ランダムな変動」として処理されることになる。

ちなみに、集団を対象とした調査やそれを担当する研究者は、聞き取りなどの調査結果において個人情報が出漏れしないことを謳い、情報提供者を匿名とするものが殆どである。こうした姿勢は個人情報の保護にもつながると同時に、研究と研究者の立場を明示している。簡潔に述べるならば、研究主体は調査の対象となる者の個人的な状況そのものを扱うために研究を行ってはいない、ということに尽きる。「データは統計的に処理されて…」といった説明は、その研究が調査対象者個人に直接反映するものではなく、調査対象者が属する性別や年齢や職業やその他の特定の状況や特徴をもつ「階層」という群についての知見として利用される、という趣旨である。つまり、統計を用いた数量的アプローチとは、集団を対象とする知見が意味をもつ自治体や官公庁や企業などの組織体や研究機関などによって主に用いられる方法であり、平均値、標準偏差などの代表値を元にして、そうした「代表的」で「一般的な」あり方に対して、いわば統括的な立場から施策（政治的あるいは業務的）や対応方法を検討し作成する際の根拠を提供するものである。

したがって、数量的アプローチとは、ある集団や階層に対しての施策や対応を行うための根拠や知見を得るために行われ、分布の代表値を把握することを主なねらいとする方法であること、そのためには調査そのものが平均値によって示されるような複数者を前提とするとともに、個々人の相違というよりももっぱら全体の平均値的な理解を求める方法となる。こうしたアプローチは、自らを厳密で客観的な方法として位置づけているけれども、実際は、平均値などの「代表値」にしても数値そのものはあくまでも統計学的抽象に基づく概念的な数量であって、

そうした数量処理過程が「厳密で客観的」であるということに過ぎない。

たとえば、複数人のデータから算出されるいわゆる「代表値」の一つ、平均値についてみると、その平均値近辺の数値をとる人数はそれほど多くはないという一般的な事実がある。グループ全体の平均身長が仮に165cm だったとして、その程度の身長、たとえば163cm ～167cm の人数が全体の半分以上といったようなことは少なく、かなり特殊な場合に限られる。また、身長の数値をいくつかの区間に分けてその中で人数が最も多い身長区間が最頻値（mode）となるにしても、最頻区間に属す人数が全体の半分以上ということも必ずしも一般的なことではない。つまり、統計的に用いられる代表値は、平均値にしても最頻値にしても、「代表値」という術語が示す程の「代表性」を示しているとは限られず、そこに分散や尖度や歪度といった統計学的な抽象を追加することによって、何とか分布の形状が想像できる程度である。

つまり、統計学とは、現象の一部についての極めて限定的な情報のみに基づいて、理論的に算出する統計的数量を提供するものであるから、あくまでも、データとして用いられた情報に基づく限定的な解釈にとどまるものである。一般的に言えば、「フィルターによって濾しとられた情報」を「ある変換によって処理」してそのように「生産された情報」という構造をもつ情報変換過程であり、いわゆる関数関係、あるいは射影関係の一つに過ぎない。そこにあるのは、情報処理過程の「客観性」であり、算出された情報の変換過程内での「実証性」であって、元々の現象や現実そのものとは直接に関係のない、統計学という抽象的演算過程における情報変換操作に過ぎない。統計的に得られた結果とは、したがって、研究対象とされている事柄の属性のうち、もっぱら分布に関連する側面についての情報分析となるのみである。

人間の社会的多様性そして実証研究の否定

ところで、統計的検定に基づいて「一般化」を行おうとする数量的アプローチの問題は、有意水準や帰無仮説といったような技術的な視点に留まるものではない。たとえば社会心理学者ガーゲン（K. J. Gergen）¹⁴⁾が、統計学的検定に基づいた従来の実証的な社会心理学に対して向ける視線には厳しいものがあり、その立場は有意性検定といった統計学的技術論への批判といったレベルをはるかに突き抜け「実証研究の否定」という全く新しい立脚点へと至るものである。ガーゲンは「…実証という言葉は存在しない。つまり、客観的事実についての理論を、実証的方法で検討するという論理実証主義的研究スタイルは棄却される…」といった極めて強い立場を主導するのである。

一見すると極端な主張にみえるガーゲンのこうした立場は、実は現代社会学の潮流を正しく受け継ぐものと考えることができよう。葛西（2005）が記しているように、社会的現実がどのように構成されるかという社会学の根本問題について、現象学的社会学を打ち立て多元的現実論を説いたシュッツ（A. Schutz）や、ガーフィンケル（H. Garfinkel）のエスノメソドロジー（ethnomethodology）による社会的現実の構成についての探求は、そうして展開の成果の一つ

として社会構成主義あるいは社会構築主義 (social constructionism, social constructivism) という立場として提起されているからである。「現実社会的に構成される」といった主張は、人と人との関係によって社会が産み出され、そのようにして社会的現実が構成されることを説く。そのため、調査対象となる人々がどのような社会的現実を生きているのかによって、価値判断や行動基準などを含めて大きな差異の存在を予想させる。そうした詳細を無視あるいは極度に抽象化して希薄化させる研究は、少数の要因によって社会現象全体を把握しようとするパーソンズ (T.Parsons) によるいわゆる「誇大理論」を経て、経験科学的な背景をもつ実証的研究による「中範囲の理論」の積み重ねによるべきだとしてマートン (K.R.Marton) によって発展的に批判されてきた。そして、その場その場における微視的な人と人との関係性の中で「社会的現実」が創り出されるという現象学的社会学を経て今日に至った社会学上の進展が、ガーゲンの主張の根底に置かれているといえよう。

社会心理学領域における実証的研究の否定というガーゲンによる主張は社会構築主義という思想的な立場を前提とするのに対して、本論文は特にそうした思想を前提とするものでない。しかし、上に示したように、統計学というアプローチは、統計学的数量処理という内部的変換過程の「客観性」(あるいは非恣意性)と、そのような統計学的算出過程の「実証性」(あるいは算出結果の一定性)という点において厳密なのであって、それによって示される統計学的結論や統計学的知見は、現実の詳細や実態についての類推の手がかりとして用いられるに過ぎないものである。したがって、有意性検定によって「実証された」とする社会心理学的知見は、そうした抽象レベルにおける、類比的な意味での「実証」なのであって、それによって人間全体についての何ほどのことが事実として「実証された」と主張できるのかは定かではない。

実際、社会心理学の領域において提起された一つの理論は、それに該当する事例と該当しない事例の狭間にあって批判的な次なる理論の提起を呼び、そうした展開が次々に繰り返されるという経過をたどることが多い。(葛西2005における「モデルの多重併存状態」を参照のこと)。そうした社会心理学的研究のうねり全体を通じて、人間の有様や行為についての理解が深まる契機となったことは確かであるにしても、一つ一つの仮説なり理論なりが「実証された」との主張がありそれらが次に「反証された」と主張されるという展開は、いわゆる「実証主義的研究」というアプローチが、研究の基本的構えにおいて現実の多様性をどれほど把握し損なっているかを明確に示しているようにも思われる。

いずれにしても、検定結果の解釈によって得られると想定されるらしい「実証」ということが厳密には不可能であるという立場においては、本論文は結果的にガーゲンの実証主義への批判と足並みを揃えることにもなるだろう。

「人間の斉一性」をめぐる心理学の歴史的陥穽

近代の科学的心理学の歴史が主に感覚についての研究から始まったということは、心理学の

発展にとってプラスの側面と同時に、心理学の方法論に関してはマイナスの側面をはらんでいたと考えられる。プラスの側面とは、「被験者であるその特定の個人による判断」という一人一人の認識や判断の個別性に注目するのではなく、いわば感覚器官の能力測定といった観点から実験研究を行えたため、「斉一性の公理」がそれなりに妥当性をもちえた点にある。つまり、日本で測定しても諸外国で測定しても「ヒト」ならば共通の身体機能・感覚機能をもつことから同様の結果となるはずである、といった暗黙の前提がそれなりに機能したと推測されることによる。そして、帰納的推論においては観測数は原理的には無限大となるところを、「斉一性」の公理が導入可能となることによって、また、統計的検定を用いることによって、観測数を大幅に低く抑えられたことになる。つまり、高々数十人から数百人程度の被験者による実験的研究によって、人類全体についての「真理」が獲得されるといった暗黙の想定が置かれていたと考えられる。

さて、感覚についての心理学的研究は当初、どの程度の強さの光まで見えるのか、どの程度に弱い音まで聞こえるか、どの程度の強さで皮膚が押されると痛みを感じるかといったような絶対閾ないし刺激閾の研究が中心であり、心理学の研究というよりもむしろ感覚機能の生理学的研究の一部とみなされるうるものだった。そうした意味では、19世紀後半、人間の弁別閾（discriminative threshold）の研究に基づいて「精神物理学 psychophysics」と呼ばれる領域が登場してきたことはいわば必然的な流れだといえよう。「ウェーバーの法則 Weber's law」は、二つの刺激の差異を弁別するために必要な刺激の相違量（丁度可知差異 jnd; just noticeable difference）に関する法則であり、それに基づくことによって、たとえば光刺激の強さに差異を感じるときには、二つの刺激の強さの物理的差異は対数関係にあるとする「フェヒナーの法則 Fechner's law」によって、精神の「物理学」が打ち立てられるに至った。物理量と感覚量のこうした関係は、後の「スティーヴンスの巾（乗）法則 Stevens' power law」に見られるように、「星の明るさを比較するとそれらの等級が主観的には二倍、三倍…と感じられるのに対して、物理的な光の強さは巾乗倍となっている」といった例によって示される内容をもつ。いずれにしても、そうした研究では、当然のように、被験者の人格的社会的文化的要因を想定することはほとんどなく、そこには「人類ならば共通の…」といった「斉一性」の思想が暗黙裏に想定されていたと考えられよう。

その後、20世紀初頭に登場したゲシュタルト心理学において、実際には動いていない対象が動いているように見える「仮現運動 apparent movement」（二つの光点を交互に点灯すると点点が移動しているように見える現象や、静止画を一定間隔で点灯することで動きを表示する映画など）の研究を通じて、視覚的認識は要素に還元されないある全体性（Gestalt）に基づいて行われているという理解へと至った。錯視の研究における「図 figure と地 ground の反転図形」や、同一の図形でありながらいくつかの異なった事物として見える「多義図形」についての研究を通じて、「感覚 sensation」の研究から、広い意味での「認識」に関わる「知覚 perception」

の心理学へと進展していった。ここでは、物理的な外的事実とそれを認識し体験する際の内的事実とのズレの存在が明確化されることによって、真に「こころ」の学問としての心理学の可能性が芽生えた時期と考えられる。

こうした研究においては、それでも「ヒト」として共通の感覚器官をもつ「人体」における体験内容の研究という意味では、やはりミルの「自然の斉一性」なる公理がそれなりに妥当しているといえるだろう。つまり、人類として共通の感覚機構と共通の情報処理機構をもつ生体を前提としているという意味での、物的な「斉一性」を仮定できるということである。そのため、ミルによれば「同様の現象がある程度繰り返されたならば、そういうこととして一般化される」のであるから、実験回数ないし被験者数をそれほど多数にすることなく、現象の一般的な把握が可能であったといえる。

しかし、「生体生理学的」な心理学から「こころ」の心理学という方向への転換に関して決定的ともいえる発見が次に巻き起こった。1940年代の知覚研究において、人間の欲求や期待や態度、過去の経験などの人格的及び社会的要因によって、知覚のあり方が強く影響されることが次々と示されたのである。たとえば、子どもはコインのサイズを実寸よりも大きく知覚しているだけではなく（重要で価値のあるものは大きく見える）、貧しい家の子どもの方が裕福な家の子どもよりもコインのサイズをかなり大きく知覚している、といった研究がその一例である。あるいは、知覚者にとって不快で忌避される刺激は知覚されにくいといった現象が「知覚的防衛 perceptual defense」として見いだされてもいた。たとえば、物事の価値観に関する単語を瞬間提示すると、本人の価値観と合致する単語は素早く認識されるのに対して、価値観に合致しない単語では反応が遅かったり、あるいは性的なタブー語を提示した場合には通常の単語以上に長く提示しなければ認識できないなどの現象がそうした一例である。後者は、精神分析における「抑圧 repression」の概念によって説明されるなど、人間の内的世界が外的事実をそのまま取り込んでいる訳ではないという点において、「こころについての心理学」の重要性を指し示す画期的な研究成果であった。

「ニュールック心理学 new look in psychology」¹⁵⁾と呼ばれるこうした潮流は、知覚心理学内部での一出来事のように捉えられているけれども、実際は心理学の方法論上の重要な転換点でもあったはずである。すなわち、「知覚のあり方が社会的要因や人格的要因によって異なる」のであれば、「人類であれば共通であるはずの感覚機能・情報処理機能としての感覚・知覚を研究する」といったようなそれまでの大前提が崩壊したと言えるからである。つまり、知覚の研究は、ニュールック心理学以降は被験者の性別、性格、職業、収入、社会的地位、宗教など多くの人格的社会的文化的要因に注目して、被験者のそうした特徴や傾向を明確にした上で研究を進めるか、あるいは、そのように異なる被験者層のどれかに偏らないような統計的な配慮を前提として行われるべき事態に至ったのである。だが、奇妙なことにニュールック心理学の登場に伴って必然となるはずの方法論上の転換は何故か全く実現されなかった。また、そ

の後しばらくして登場する認知心理学（cognitive psychology）にしても、そうした方法論上の切り替えは行われことなく、やはり「人間の斉一性」公理のレベルでの研究に留まっていた。そうした方法論上の転回が実現されなかった事情は定かではないが、感覚知覚の研究者たちはニュールック派の登場によって従来までの方法論による研究が不可能になったわけではなかったことが大きく影響していたと思われる。たとえば、科学におけるパラダイム・シフト（paradigm shift）を唱えたクーン（T. Kuhn）¹⁶⁾が示しているように、天体の運行に関してかつての天動説による計算式がひどく煩雑であったにしてもそれなりに予測に使える以上、当初は特に地動説に切り替えるべき原理上の切迫性がなかったといった状況と類比されるのかもしれない。

いずれにしても、ニュールック派による心理学上の功績は、文化的社会的な要因による個人差の存在を明確に描き出し「人間の斉一性」的な想定を根底から覆したことだったといえる。したがってそのように多種多様な被験者側の特性を考慮するならば、実験や調査では研究対象となる被験者層を網羅するなどの配慮を迫られるため、被験者数や観測度数は必然的に大きな数値とならざるを得ない。そうした大規模な研究によって初めて「人類」についての知覚研究となるという理解がその当時に生まれていれば、その後の心理学は全く異なる発展の道筋をたどったと思われる。なお、ニュールック派による発見の意義が見落とされたもう一つの理由は、統計的な研究とは母集団から「無作為抽出された標本」に基づいて行われるという要請が、当時の心理学の世界において無視されていたか、あるいは不当に過小評価されていたことにあると推測される。仮に「無作為抽出」という要請に対して心理学の世界が厳格な態度で向かっていたならば、たまたま構成した実験群の被験者たちが、偶然にもある特定の文化的社会的特徴を共通に持っていたなどの事実上の「系統性」に対しても敏感であっただろうし、そうした偏りによって調査や実験研究そのものが台無しになる可能性を抑えるという、方法論上のさらなる展開が得られていたはずだからである。つまり、「人間の斉一性」といった想定が暗黙のうちに瀰漫することによって、統計学的な観点からいえば「標本の無作為抽出性への無関心」という違背が生じ、また、研究面からいえば「文化的社会的階層に関する被験者の偏りについての無関心」といった重大な違背が心理学的研究の根底に組み込まれていたことが推測されるのである。

さて、こうした感覚知覚研究とは全く別個に、心理学は当時発展を続けていた生理学的研究からも強い影響を受けることとなった。条件反射学を唱え、後に「古典的条件付け classical conditioning」ないし「レスポナント条件付け」と呼ばれる行動メカニズムを明らかにしたロシアの生理学者パブロフ（I.P. Pavlov, 1849-1936）の研究は、ワトソン（J.B. Watson, 1878-1958）によって、自然科学と同一の方法論を有するアプローチとして高く評価され、行動主義（behaviorism）と呼ばれる立場として発展することとなった。内省（introspection）によって「こころ」の内界を探るという当時の「意識心理学 consciousness psychology」を批判したワ

トソンは、客観的に把握可能である「行動」のみを心理学の対象とすべきだと提唱したのである。次いで、「オペラント条件付け operant conditioning」を発見し研究を展開したスキナー (B.F.Skinner, 1904-1990) によって、そうした客観主義は「新行動主義」あるいは「徹底的行動主義」としてさらに純化され「学習心理学 psychology of learning」として発展していった。そのようにして心理学に君臨する勢いを得た行動主義という立場は、たとえば対象となるネズミなどの実験動物の種固有の感覚知覚システム、情報処理システム、行動パターンを基本として研究を進めていることから、ミルの「斉一性の公理」に合致するような身体器官的レベルでの物理的生理的同一性を前提としたアプローチといえる。行動主義はそれと全く同様の発想を人間についても適用しようとしたものであった。このように、モノとしての生体とその機能に関する客観性に魅了された行動主義の心理学は、「個人的な体験内容」として構成されると思われる「こころ」の学問としての立場を完全に排除し、一種の自己矛盾と見なされるべき「こころに一切言及することのない〈心理学〉」を形成したのである。

ところで、当初は行動主義的な実験心理学者として研究活動を開始したマスロー (A.Maslow, 1908-1970)¹⁷⁾ は、当然ながら、ネズミと人間の動機の構成が異なることに気づき、自己実現欲求を含む高次欲求の存在を提起し五段階からなる動機理論としてまとめていった。そうした研究の観点は、心理学の第五勢力と呼ばれる「人間性心理学 humanistic psychology」が展開していく一つの起点ともなった。なお、マスローの動機理論には動機の階層性が想定されていて、下位の動機が充足されるまでその上位の動機は解発されないという仮定をおくものである。マスローによる動機理論では、下位の動機は他の動物にも認められるのに対して、高次の動機はほ乳類などに共通のものを含み、最高次の第五段階の動機は人間独自のものとなっている。ちなみに、最下位に位置する動機は「生理的欲求 physical need」といい生命維持と生存に必須となるあらゆる欲求が含まれる。それがある程度充足されることによって、一つ上の動機「安全欲求 safety need」が現れる。これは、心身が損なわれないことと共に、世界の安定性と予測可能性、制御可能性の獲得を含む動機とされる。この二つは、生体・動物としての存在に必須であるという点において「基本的欲求」とも呼ばれる。さて、三番目の動機は「帰属と愛への欲求 belongingness and love need」といい、家族・職場・友人などとの人間的で心の通い合う関係を求める動機である。この動機がそれなりに充足されると四番目の動機「自尊欲求 self-esteem need」、自らの存在とその意味や価値が他者に承認されたいという動機が現れる。この三番目と四番目の動機はその性質から「社会的動機」とも呼ばれる。こうした動機レベルまでに至るならば、さしあたり、生命体として安寧に存在し、暖かい社会的な交流の中で自らの価値が承認されているという極めて望ましい状態にあると考えられる。しかし、マスローの動機理論の最大の特徴は、その上にさらに「自己実現欲求 self actualization need」という高次のレベルを明確に位置づけたことにある。つまり、自らの存在の意義や意味、生きている証を求めていくという人間は、他の動物と比較するならば、欲求を充足し尽くすことのない存在で

あること、したがって極めて充足した望ましい状態さえも自らの更なる自己実現のためには打ち捨てさえするという一種の自己破壊的な選択の可能性を含んでいるとされた。ここにおいて、人間は環境や状況に隷属しそうした要素によって予測し尽くされるのではなく、高度に柔軟で予想できない程までに多様な行動を取り得ることが明確に指摘されたといえる。

こうした立場は、種としての生体上の同一性を前提として、平均的な意味合いでの人間の行動予測性を謳う行動主義の立場とは真っ正面から衝突する。しかし、生体という実体を前提として測定や調査によってデータを得て、統計的検定によって研究結果の妥当性を主張する「客観的」な行動主義や数量的アプローチの圧力には勝てず、マスローの動機理論や人間性心理学的なアプローチそのものが心理学の主流となることはなかった。また、社会学領域においては大きな影響を与え学問的発展の基軸ともなった現象学的アプローチは、ヨーロッパの心理学研究においては一定の展開を見ていたにもかかわらずアメリカではほとんど顧みられることなく推移したし、現象学そのものへの国内での関心の高さにも関わらず、心理学領域では長年花開くこともなかったといえる。

なお、現代に至っては、医療における高度な生体測定情報処理システムによって脳内の血流パターンや脳波パターンの把握、あるいは「楽しい」といった主観的体験と微少なホルモン量などの変化の解析などによって、人間の「思い」という主観的な体験が生体上の実態として徐々に把握可能となってきている。あるいは心身医学領域においては、笑いと免疫との関係が明確に取り出されたり、腫瘍と「思い」の関係を医学的に扱う精神腫瘍学（psycho-oncology）が展開されるなど、いわゆる主観的なあり方が一定の生体的データとして把握できる段階によく到達したといえる。情報解析に基づくこうした新たな数量的アプローチの展開は、少なくとも、「主観性」という事柄の意義を、測定技術的にそれなりに正当に評価できる段階に至ったことを示すものであろう。

さて、行動主義以降のその後の展開についてみると、感覚と知覚の心理学そして条件付けをその内容とする学習心理学は、ともに、人体の器官的な実体を前提とするという意味での「客観的な」研究として位置づけられることによって、心理学は極めて偏った形での「科学的で客観的な」学問としての地位を固めるに至ったといえる。同時に、そのための方法論は、実験や調査によって数量的データを収集すること、それに基づいて統計的仮説検定を行うこと、という形式にほぼ統一されていった。その際、すでに述べたニュールック心理学における発見、すなわち、人間の知覚は文化的社会的要因によって大きく左右されるという事実から容易に推測されるはずの、個々人の認知や行動の個別性や多様性といった側面はいわゆる「科学的心理学」においてはほとんど無視された。そうした中で神格化と呼べる程にまで絶対視されたのが「統計的仮説検定」であり「有意水準」値であった。そこでは知覚ならびに行動の本質的多様性という認識はあらかじめ排除されているために、人間の知覚や行動には、本来、「真の」中心的な数値が存在しているという見方に基づくこととなる。次いで、そうした「真の」数値をかき

乱すような「攪乱」要因が作用することによって誤差が生じるために、実測値は理想的な「真の」数値からは逸脱し、ある平均の周りに分散して「分布する（たとえば正規分布として）」とされた。つまり、人間の一人一人の多様性は非本質的であり、平均値などの代表値からのズレは単なる測定上ないし現象上の「誤差」に過ぎないものとして扱われていたといえる。そうした姿勢の根底にある学問的な関心事とは、もっぱら人間の行動における「一般的な傾向」であって、個別的で独特な一人一人の行動の多様性ではなかった。そうした流れが現代の心理学の基礎を形作っていったため、一人一人異なる他者をどのように感じ取り理解するかといった訓練が心理学の基礎教程とされることは稀であるのに対して、統計学を中心とした教育が例えば心理学の研究法という科目として整備されるという展開をたどったのである。

母集団の不定性ということ

人間の一般的傾向を調べるという意味では、上に述べてきた感覚知覚研究および行動主義的研究は、いわば平均値至上主義とでも名付けられる方向にあるといえる。さて、平均値とは何らかのグループについての一つの代表値であるから、標本を抽出するための母体となるグループ、一般に「母集団」が何であるかを研究者は明確に設定せねばならない。統計学では、全数調査のように必ずしも母集団全体を調べ尽くしうる訳ではないことから、母集団から標本を抽出しその標本の平均値や分散などの数値に基づいて母集団の特性である「母数 population parameter」を推計するのである。

ところで、感覚知覚研究および行動主義的研究には「母集団」という概念はほとんど登場してこない。母集団から取り出された n 個の標本についての統計的分析から母集団の特性値たる母数が推測されるにも関わらず、それらの研究はおおむね母集団についての明確な指定や記述を欠いたまま行われている。その理由の一つは、精神物理学や行動主義が暗黙のうちに研究対象としているのがいわばヒトという動物種の生体的器官の特性であるからといえる。あえて述べるならば、それらのアプローチの母集団とは暗黙のうちに人類という種全体を意味するから特に言及する必要もなかったと解釈される。

こうした視点にたどり着くことによって、統計的アプローチに基づく心理学研究は、極めて簡潔に二つのグループに分けることができる。一つは「母集団についての記述がない研究」、他方は「母集団についての記述がある研究」である。母集団についての記述がない研究は、基本的に「ヒトの斉一性」を前提としてヒトの一般的特性としての器官の側面を前提として、個々人の反応の相違そのものを研究対象とはしないことが考えられるのに対して、母集団についての記述がある後者は、様々な要因によって「個々人は相互に異なる」ことを前提としているために、被験者属性と階層性（年齢、性別、職業など）を意識した研究となる。たとえば、社会調査と呼ばれるような比較的規模の大きい研究では、年齢構成、性別、職業、地域などの要素をあらかじめ考慮に入れ、ある信頼区間で母数を推定するためにはどの程度の標本数を必要と

するかといった調査研究の設計を行うことになる。仮に、男女によってある特性値が異なることを信頼区間95%で調べることにして、予備調査では男女それぞれの特性値がおおむね分かっている場合は、そうした数値を計算に入れつつ必要な標本数を算出するという具合である。その際、得られた結果はあくまでもその標本に関しての数値であって、それらの数値は、たとえばある地域やある特定の状況の人々を母集団と設定した上で、その全数調査の代わりに、標本抽出によって母集団についての推計を行うという明確な方法論に基づいて行われる。そのため、それらの標本特性値は設定された母集団特性を推測するために使われるのであって、そうした限定的な結果をそのまま人類全体にあてはめて一般化することはない。研究結果と研究対象は共に、あらかじめ設定された母集団という枠組みの中に留まり、その中での結果の解釈とその母集団について推計となるからである。

こうした研究手法と比較してみるならば、心理学研究誌に投稿されている実験研究や比較的小規模な統計的調査研究は、被験者数や調査回答者数が単に少ないだけではなく、被験者や回答者の属性もそれほどに層別化されずに行われるものがほとんどである。実験研究では高々数十名程度の被験者数に留まるものが多いし、調査研究の場合でも回答者数が数百名に及ぶものは多くはなく、また、回答者層もある特定の年齢や所属（教育機関や職場など）に偏るなど相当に縁故的であって、標本抽出の基本原則である無作為抽出標本ではないこともそれほど例外的ではない。さらに、そうした研究のほとんどが、観測の対象とされた被験者や回答者についての簡単な記述があるのみで「母集団についての記述のない」研究である。つまり、そうした標本集団から得られた統計的数値が一体どういう母集団における母数の推定に用いられるのかが明示されていないのである。そうした実態にも関わらず、論文の記述や結論は、研究に協力した回答者や被験者の所属や社会的階層などの実際の枠組みを超えて、あたかも国民や人類全体がこれこれの傾向や特徴をもつと主張しているように読めるものも少なくない。自らの研究がどのような一群の人々を母集団として設定しているかを明確に記述することがないため、標本から得られた結果が誰の何についてのものなのか、そうした枠組みや適用範囲やその限界について、研究者自身が明確に認識していないためであるといえる。なお、これと類似した指摘としては、有意性検定の結果からの推論が、回答者群という「集合 the aggregate」に対して限定的に行われるべきにも関わらず、それとは全く異なる「一般 the general」への推論過程と混同されているという指摘もなされている¹⁸⁾。

こうした問題を次のような例で具体的に掘り下げてみよう。（なお、質問項目の内容がヒトとしての生体的同一性に関連するようなものは除外する。）

例えば、A大学にはX、Y、Z学部があるとして、X学部二年生100人のうち50人の男女学生についてアンケート調査をした場合を考えてみる。この50人の男女学生から得られた或る特性についての平均値や比率などの結果は、さしあたりはその50人についての統計的事実である。したがって、その数値をそのまま提示することは統計的推論や検定以前の単なる事実の提示と

なる。次に、標本値にはたとえば性別ごとに何らかの傾向があるか否かについて統計的検定、例えば男女別の群における平均値の差の比較を考え、t 検定などの（パラメトリックな parametric）検定を行ったとしよう。その場合、調査によって得られた標本平均値、標本分散などは、何らかの母集団の特性を推計するために用いられるのであるが、さて、その場合の母集団とは例えば次のうちのどれであろうか。

「X学部二年生100人のうちから50人を調べること」によって—

- 1) X学部で調べた男女50人がそのまま母集団で、彼ら50人の特徴を知りたかった。
- 2) 本当はX学部に所属する同学年100人全員についての特徴を知りたかった。
- 3) 本当はX学部に所属する上級生・下級生を含む全員についての特徴を知りたかった。
- 4) 本当はX学部、Y学部、Z学部からなるA大学の学生数千人ないし数万人の特徴を知りたかった。
- 5) 本当はA大学とその近郊にあるB大学の学生、数千人から数万人の学生の特徴を知りたかった。
- 6) 本当はA大学のあるC地方に住む数十万ないし数百万人の住民の特徴を知りたかった。
- 7) 本当はA大学のあるD国全体の国民の特徴について知りたかった。
- 8) 本当はA大学のあるアジア地域の住民全体の特徴について知りたかった。
- 9) 本当はA大学のある地球上の人類全員の特征について知りたかった。

などなどといったような様々な「母集団」が想像される。ここでは年齢条件などを含まない例を示したが、そうした要因を入れた上で想定される「母集団」の数と種類は一挙にふくれあがってしまう。さて、仮に有意水準5%なり1%なりにおいて、性別ごとに相違があるということが統計的に有意だと判定されたとして、それではその結果がどの範囲にまで拡大して解釈可能なのだろうか。すなわち、「ある国」の「ある地域」の「ある大学」の「何年生かの学生」を回答者として取られたアンケートにおける有意な結果は、上級生や下級生にも適用されるのか、隣の大学の学生にも適用できるのか、別の都市の大学生やよその国の大学生にも妥当するのか、あるいはあらゆる社会階層や人類全体にも拡大されるのだろうか—。

常識的には上のリストの(1)(2)あたりが想定される程度であり、高々(3)(4)あたりが想定の上の限界と思われる。たとえば(1)を母集団とするならば、「数ヶ月前や数ヶ月後にはそれなりに違っているかもしれないけれども、さしあたり現在の特定の時点でのデータを標本として、通年での状況を母数として把握したかった」といった説明が唱えられるかもしれない。また、(2)を母集団とするならば、「X学部二年生100人全員が母集団であるけれども、全数調査が難しいので、そのうちの50人だけを標本抽出してみた」といった社会調査研究の視点が唱えられるかもしれない。(3)や(4)を母集団とするのは普通はやや無理があると判断される可能性があるにしても、

社会調査研究の視点からはそれなりに納得させられてしまうかもしれないし、その反対に、現場を熟知している心理学者からは「学生の頃の一年の差は大きな違いを生むし、学生の質は学科ごとにより違う」という批判が出ることも予想される。もちろん、(5)～(9)は、統計技術的にはおおむね無理だと判断されるにしても（サンプルサイズや層別化の設定の問題など）、母集団についての記述を欠く心理学的研究の多くは、そうした母集団設定を明示しないことによって、暗黙のうちにそうした拡大解釈や一般化を行ってしまう可能性をあながち否定できないであろう。なお、そうした指摘の根拠の一つとして、正規分布とそれを基本とするパラメトリックな検定法が心理学研究では主に用いられていて、ノンパラメトリックな(non-parametric)検定法は一般的ではないという事実が挙げられる。たとえば、標本データの順位に基づくノンパラメトリックな順位和検定では、標本数そのものがいわば「母集団」サイズとして計算されるから本質的に「(回答者) 集合 the aggregate」についての検定手法と考えられる。それに対して、パラメトリックな検定法は、そういった限定的な標本についての検定ではなく、膨大な量の標本を前提とする「大数の（強）法則」が示すような意味での「一般 the general」についての検定となるからである。

ところで、「地域住民」や「国民」といった概念が「学生」という概念の上位概念、つまり「学生」といった概念をその傘下に納める概念、とされたとしても（厳密にはそうした設定が妥当な状況か否かを吟味する必要がある）。すると、「地域住民」概念の下には「学生」と同時に「非学生」である地域住民のカテゴリーが想定されることになり、「地域住民」で「非学生」である「サラリーマン」「主婦」など様々な社会的階層のカテゴリーがその下に置かれるだろう。さて、「学生」を回答者として得られた結果をたとえば「地域住民」についての調査だとして母集団を設定したならば、「地域住民」概念の下にある「非学生」たる社会的階層の反応を、自動的に「学生」による反応と同じであると見なしたことになる。これは、「人間の斉一性」ならぬ「学生と非学生の斉一性」仮定を暗黙のうちに果たしたことを意味する。学生と非学生の反応が同じであることを積極的に肯定するようなテーマについてはともかく、基本的には唐突な想定といわざるを得ない。そうした仮定を置くくらいならば、回答者を学生に限定して調査せずに、最初から様々な社会的階層から無作為抽出（層別化などの配慮を加えた上で）する調査方法を採用すべきとされるだろう。このように、ある限定的な社会的ラベルを付与された標本集団において得られた標本統計値を、その社会的ラベルとは排他的な社会的階層にそのまま拡大適用することには問題があることから、X学部二年生のうちの50人の「学生」についての調査研究が、(1)から(9)までの母集団について、どのレベルまでが母集団として想定可能かといえば、厳密には「(1)X学部で調べた男女50人がそのまま母集団」、そして、「(2)X学部に所属する同学年100人全員が母集団」といった想定までが許容されることとなるだろう。もしもそれを仮に(6)から(9)までの母集団として想定したならば、調査を行った「学生」集団において得られたに過ぎない標本値を、「非学生」である多種多様な下位集団全てにおいて採用するとい

う極めて乱暴な想定をしたこととなる。

なお、標本と母集団との関係についてのこうした議論は、実はそれに先だって「母集団から標本を無作為に抽出する」という大前提の元で進められなければならないことである。しかし、現状としては心理学的研究の多くが、調査研究の回答者あるいは実験研究の被験者として、特定の学校や組織に所属する者を系統的に用いており、無作為抽出 (random sampling) にはなっていない。そうした「縁故的」標本設定によって得られた標本値は、当然ながら、回答者などの属性である「学生」「サラリーマン」などの社会的ラベルや地域性などの属性によって特徴づけられるため、その適用が極めて限定される数値とならざるを得ない。したがって、心理学的研究の多くは本来、限定的な標本集団と母集団についての研究なのであって、そうした研究結果をそれ以上に拡大適用ないしは拡大解釈できないという統計学的制約のもとに行われていたはずである。こうした吟味に基づいて、本論文では以下の要請を行う一。

要請2 統計的有意性検定を用いる場合、調査研究における回答者あるいは実験研究における被験者について、彼らの属性について明示せよ。そして、有意性検定はどのような母集団を想定して行われるかを明示せよ。

ここまで「要請1」「要請2」を示してきたが、アンケートや質問紙による調査研究などがある有意性検定を用いる場合、詳細に示すならばおおむね次のような項目が考慮され記述されねばならないだろう。

・調査対象者の属性について

- a) 回答者についての層別化は行っているか否か。
- b) 層別化をしていないとすればそれはなぜか。
- c) 回答者の抽出は無作為抽出か、何らかの系統的ないし縁故的抽出か。
- d) 一群の回答者の属性についてのラベルが、それら一群の回答者の状況を確かに捉えているか否か。例外的状況は皆無か。あるとすればどの程度の割合なのか。

・母集団についての記述の有無

- e) 標本とされた回答者群から得られた標本値は、どのような母集団についての母数を推定するために用いられるのかが明記されているか。

・有意性検定に関する三つの技術的問題

- f) 母数を推定する統計的有意性検定において、有意水準はどの程度に設定されているか。また、そうした数値を採用する理由は何か。

g) 有意性検定において、対立仮説を誤って採用しないでいる「第二種の過誤」について、どのような配慮を行っているのか。

h) 有意性検定において、第一種、第二種の過誤との関係において、サンプルサイズは適切に設定されているのか否か。

・検定結果解釈の局限化

i) 有意性検定の結果は、想定された母集団についての限定的な推計として厳密に解釈されているか否か。

すでに述べたように、大規模な社会調査的な研究の場合は、おおむね何らかの層別化を行い、得られた標本統計量から推計されるべき母集団の設定が明確であり、推計された母数の解釈も適切な枠組みの中で行われることが多い。したがって、(a)～(i)までの10項目についての記述を求められたとしても、ある程度まで対応可能と考えられる。それとは対照的に、比較的小規模な調査研究や実験的研究の場合は、おおむね「母集団」についての想定を欠くことが多く、上記リストのほとんどについて記述がないだけでなく、そうした吟味そのものを行っていないなど問題の根は深いといえる。そうした方法上の限界を抱えた数量的研究が特に指摘されることなく長年刊行されてきたことについて、以下では少し異なる角度から眺めてみることにする。

Ⅲ. 検定結果の解釈をめぐる問題へ

「ラベル」の意味の不定性

社会科学から目を転じてみると、自然科学における物理学や化学などの結果は「自然の斉一性」公理によれば直ちに一般化され、一定の条件の下では他の研究所でもよその国でもそうだと主張されることを許容することによって、あまねく通用する結果として扱われるとされる。しかし、実際には物理学や化学などの領域においては特に「斉一性」の仮定を必要としている訳でもない。物理学における現象はすでに理論と実験によって確認されている膨大な数量の素粒子や原子や分子から成り立っている以上、標本として取り出されたものに起きたことをそのまま母集団に起きたこととして扱うか、状況によっては母集団という概念そのものを必要としないほどである。あるいは、生体を対象とした研究もその対象が細胞やDNAを用いて試験管の中での生化学的反応に基づく場合などは、一回の実験によってたとえば $10^6 \sim 10^{10}$ 程度の分子レベルでの反応として変化が起きるといわれ、一回の実験で地球上の全人口を対象にして調査を行った観測数をはるかに超えるものとなる。また、生物学や医学では、こうした分子レベルの研究はともかくとして、生体器官や個体を対象とする場合はそこまで圧倒的な数量となることはないにしても、同一の種といった生物学的同一性を根拠として、研究対象である生体器官や被験体が基本的に等質であることを前提とすることができる。そして、研究対象となって

いる生体や被験体の物質的現象についての実験的研究が、その現象が生起している等質的な物質的実体の測定や観測によって解明が行われるのであるから、観測している現象と研究している対象との間にはズレはない。したがって、実体的な生物学的同一性ないし等質性に基づいて「斉一性」公理が自然に該当することとなるだろう。

これに対して、前章に挙げたX学部に所属する「学生」についての調査研究のように、標本抽出の前提とされるべき母集団が必ずしも一意に決まるとは限られないとき、心理学という学問領域が自然科学や医学などとは研究方法論に大きな違いをもつことに当然ながら気がつかざるを得ない。つまり、調査研究の場合、ヒトという種であれば基本的に同じ反応をするのか否かという趣旨での研究ではなく、生物学的な実体とは直接に関係のない「学生」という概念、あるいはそうした社会的ラベルによって同一グループと見なされうる回答者たちの反応が研究対象となっているからなのである。つまり、母集団を特定化する上で問題となっていることは、回答者の生物学的同一性などのように実体が存在するものではなく、たとえば「学生」といった社会的ラベルによってグループ化された回答者が、それ以外の社会的ラベルである「若者」「二年生」「所属学部」「所属大学」「地域住民」「国民」「人類」などなどのラベルとの間でどのような関係になっているのかという認識上のことなのである。

しかし、その際に問題になるのはラベルの定義には現実に困難がつきまとうことであり、あえて一般意味論を持ち出すまでもなく、ラベルによって示される集団とその実態との間には必ずしも厳密な対応関係がないという点である。たとえば、学生が「若者」であるという必然性がないだけではなく、「二年生」というラベルには年齢的にも在籍年数上も多様性があり、「所属学部」にしても編入生なり転部生なり他大学などからの短期留学の学生も含まれているなど、回答者層につけられた「学生」というラベルによって回答者の等質性を厳密に主張することには困難が伴う。つまり、「学生」を回答者として行った調査の母集団は何かという問いかけは、実はそうした議論に先だって、標本となったグループのラベルとその実態とが必ずしも一致してしないという本質的な問題を引き出すことになる。これは「地図は現地ではない Map is not territory」すなわち、言葉などで抽象化された事柄は現実の対象物と一致するとは限らないという、一般意味論による基本的指摘と同一の事態である。

言葉のもつ本来的な抽象性のゆえに、ラベルとラベルによって指し示される実態との間には何らかのズレが存在していることは極めて原理的な問題である。したがって、このことに何らかの態度表明や説明をせずに済ませてしまうことは研究としては不誠実と考えられるのに対して、心理学を含む社会科学系の研究論文においてこの点についての克明な説明を含むものは極めて限られている。その理由の一つは、テーマそのものがそうした「概念」に関する研究である場合を除くならば、一般意味論的な原理的議論を行うだけの余裕が論文執筆に際してテーマ的にもスペース的にも許される状況には無いということであろう。ラベルと実態との混同に関わるこうした問題についての無知なり感性の低さなりが原因であったり、あるいは論文誌上の

余裕の有無といった紙面上ないし編集上の事情であるにしても、いずれにも共通していることは「そうした本質的で原理的な問題について記述がない」という事実であるから、こうした状況を「無記」という言葉で表すことにする。

本質的問題についての「無記」とその帰結

この「無記」という側面から、統計的手法を用いた心理学的研究などの論文における実態を捉えてみるならば、次のような特徴を指摘することができる。それは、上に示したような、「標本のラベルについての無記」、そして、有意性検定をめぐる技術的問題がそれに由来するところの「母集団についての無記」、という二つの「無記」がみられることである。統計的検定に際してその基本事項に触れないでいることは、そうした「無記」という状態にあるということだけではなく、そうした本質的な問題に関連する事項についての記述が削除されたりあるいは改変されるといったような実際上の変化を促す。たとえば、「標本のラベルについての無記」についてみるならば、回答者である「学生」といったラベルによって指し示される人々が現実的には年齢や経験や職業等々において様々に異なっているという事実が、「学生とは、学生として大学などに属する者」といったようなトートロジックで事務的なラベルによってその実質を抜き去られるということであったり、あるいは、「学生は…」と書かれるべき表現そのものが、正当な理由もなく「回答者たちは…」「被験者は…」から「人は…」といったように抽象度の高い表現に言い換えられるなどである。

また、これまでに詳細に議論してきた母集団の設定についても、そうした記述が全く示されない「無記」によって、母集団をいずれかのレベルに設定することに伴う様々な問題が覆い隠され潜在化されてしまう。統計的有意性検定の技術的な問題が見過ごされてきた背景には、長年の間、「母集団についての無記」という事実があったことを本論文は指摘するとともに、上述の二つの事柄についての「無記」によって、表面的には何ら本質的な問題が存在しないかのように、多くの論文は当該研究テーマについての記述のみを淡々と進めてきたことを指摘するものである。

さて、二つの「無記」によって本質的な問題を潜在化させてしまうならば、そこにはいわば「肯定的」に評価され得る側面と極めて否定的に評価される側面とが付随することになる。「肯定的」な側面の一つは、これによって、原理的で困難な問題にいつまでも関わり煩うことがなくなるという研究上の利点が生じることである。すなわち、研究の内容をもっぱら実験条件や被験者条件を多様に組み合わせたり追加するといった技術的側面に当てることが可能となり、実験的研究を系統的に生産できるようになること、あるいは、経済的時間的な制限によって小規模と成らざるを得ない調査研究も、その本質的問題点を特に考慮したりすることなく研究テーマそのものに即して調査研究を生産できること、である。つまり、「二種類の無記」によって本質論を避けることが可能となり、逆説的な意味において、それぞれの研究テーマに関する

研究としての「実績」を積み重ねることが容易となる。さらに、そうした論文執筆の形態が常態化するならば、「無記」とされた問題は扱わないことが関連領域の研究者の「不文律」であるかのように社会的慣習として沈殿する。こうした点は、「心理学研究についての社会学」といった新たな観点からの追求を生むだろう。というのは、ここまでの議論は、統計学的に本質的な問題を回避してしまうことによって、有意性検定を前提とする研究自体の妥当性の低下を憂慮する立場から行われてきたのに対して、心理学的研究は長い歳月の間、そうした問題の解消に向かうことがなかったのだから、結果的には「研究についての経営的判断」を優先して、「有意性検定をめぐる本質的問題」をいわば犠牲にして「当該研究領域における研究の量的な展開」を図るといった損得交換 (trade-off) を研究の根底に置いたと考えられるからである。心理学研究の多くは、現実にもそうした喩えによって把握される状態にあるといっても過言ではないであろう。

さて、「無記」に伴う否定的な側面は、本論文におけるここまでの議論全体を通じてその内容はすでに明らかにされてきている。結論的に述べるならば、統計学的手法においてどれほどの厳密さを追求しようとも、得られた結果の解釈段階において本質的な困難を抱えている以上、研究そのものが砂上の楼閣でしかないということである。すなわち、得られた標本値がどのような母集団についての推定となるかを明確に提示できないという限界のために、得られた結果の解釈を厳密に展開できない場合、最終的には何らかの例え話である「比喩的な解釈」とならざるを得ないからである。その一例としては、(1)～(9)の母集団の設定に関する議論において示したように、ある属性の人々から得られたデータを分析しながら、調査結果をその人々の記述に留めるのではなく、それとは異なった属性をもつ集団や人々にまで適用範囲を拡大して解釈してしまうことなどが挙げられる。

また、「無記」によって生じうるもう一つの否定的側面としては、本質的な問題や原理的な問題についての言及をせずにいることが研究上の慣行として進展するほどに、研究理念の退嬰化や些末化の危険が増すことであろう。統計的手法に関わる本質的な問題については「無記」である以上、研究テーマの細分化によって些末な事柄に拘泥することであったり、研究テーマそのものが研究方法や分析方法あるいはそれらの技術的側面などへの偏向であったりすることによって、人間の「こころ」についての実質的知見が積み上げられ深められる契機を失うことにもなりかねないことであろう。

以上、本論文は、統計的有意性検定を用いる心理学的研究の問題点について一定の解明を行い、その適切な運用に向けた二つの要請を示してきた。心理学における統計的アプローチがこうした指摘を明確に乗り越えられない場合、得られた結果の解釈に際して混乱や破綻を来していると考えられるため、従来、質的心理学的アプローチに対して方法論上の指摘や批判を行ってきた数量的心理学的アプローチではあるが、それに見合うだけの適格性を持ち得ないと判断されるべきものとなるだろう。

文 献

- 1) 葛西俊治「解釈的心理学研究における理論的基盤とアブダクションに基づくモデル構成法」札幌学院大学人文学会紀要 第78号, pp.1-26, 2005
- 2) J. ミル『論理学大系第三巻 論証と帰納』大関将一・小林篤郎訳, 春秋社, p.44, 1958
- 3) P.E.Meehl “The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Qualify Accuracy of Risky Numerical Predictions” in “What If There were No Significance Tests?” L.L. Harlow, S.A.Mulaik, & J.H.Steiger (Eds), Mahwah, NJ: Erlbaum, pp.393-425, 1997
- 4) D.H.Johnson “The Insignificance of Statistical Significance Testing” Journal of Wildlife Management 63(3) : pp.763-772, 1999 James town, ND : Northern Prairie Wildlife Research Center Online. (<http://www.npwr.usgs.gov/resource/methods/statsig/statsig.htm>)
- 5) D.E.Morrison & R.E.Henkel (Eds.) “The Significance Test Congtroversy”, Aldine, 1970 (内海他訳『統計的検定は有効かー有意性検定論争ー』梓出版, 1980年)
- 6) P.E.Meehl “Theory-testing in psychology and physics : A methodological paradox”, Philosophy of Science, 34, pp.103-115, 1967 (Reprinted in “The significance test controversy”, pp.252-266, 1970)
- 7) 伊勢田哲治『疑似科学と科学の哲学』名古屋大学出版会, pp.35-40, 2003
- 8) J. コーエン『現代心理学展開 2 心理的確率』北村晴朗・佐藤伶訳, 誠信書房, p.34, 1976
- 9) 『原子力発電の危険性 調査, 資料, 理論, そして闘い』技術と人間社1976
- 10) 葛西俊治「低確率事象の認知研究」北海道工業大学研究紀要 第17号, pp.263-271, 1989
- 11) 永田靖『サンプルサイズの決め方』朝倉書店, p.8, 2003
- 12) 山内光哉『心理・教育のための統計法』(第二版)サイエンス社, pp.114-115, 1998
- 13) C. クームス他著『数理心理学序説』小野茂監訳, 新曜社, pp.176-185, 1974
- 14) K. ガーゲン『もう一つの社会心理学：社会行動学の転換に向けて』杉万俊夫他監訳, ナカニシヤ出版, 1998 (“Toward transformation in social knowledge”, 1994)
- 15) 和田陽平他編『感覚+知覚心理学ハンドブック』誠信書房, pp.131-134, 1969
- 16) T. クーン『科学革命の構造』中山茂訳, みすず書房, 1971
- 17) A. マスロー『人間性の心理学』小口忠彦監訳, 産業能率大学出版部, 1971 (A. H.Maslow “Motivation and Personality”, 1954)
- 18) D. Bakan “The Test of Significance in Psychological Research”, in “The Significance Test Congtroversy” pp.244-245, 1970

Application Limits of Statistical Significance Test in Psychological Studies

KASAI Toshiharu

In the process of confirming the interpretative and qualitative approach in psychology, it turned out that the statistical metrical approach needed a thorough re-examination because it has been questioned for years about its validity and applicability to social sciences. Although metrical psychological researches and studies have been published in professional journals only when they succeeded in rejecting the null hypothesis on the significance level of 5% or 1%, the conventional statistical way of this test was criticized technically, methodologically, and epistemologically by a classical book titled “The significance test controversy” (1970). However, no improvements have been made about its misuse, misinterpretation, and meaningless abuse until now as described in “The insignificance of statistical significance testing” (1999).

The present paper looked into two basic statistical issues, 1) how to set the significance level, 2) what kind of the statistical population to be supposed from which samples are drawn, and found that “the sacred level of significance, 5% or 1%” should be modified according to the purpose of each study, and that the population should be differentiated and defined strictly in order to avoid the implicit supposition of “uniformity of man”.

The history of psychology was viewed as a movement toward the general understanding of human nature by neglecting its personal, social, cultural and historical multiplicity and also toward a so-called scientific discipline such as the psychology of perception and behavioral sciences by employing the statistical significance test. However, since the technical requests for clear definition of the attributes of a) sampled subjects and b) its population are not easily satisfied as General Semantics indicates, it was pointed out that the interpretation of significance test results does not necessarily yield a generalized conclusion for a “hard” science. Discussion was made about the phenomenon of “non-description” about statistically fundamental problems in statistics-oriented psychological papers.

Keywords: interpretative and qualitative psychology, metrical psychology, statistical test, significance level, population, sampling, uniformity of man, General Semantics, non-description

(かさい としはる 本学人文学部教授 臨床心理学科所属)