

心理学のデータ解析における統計法上の仮定について

— t 検定の場合を中心に —

石 田 潤

心理学の研究においては実験や調査が多用される。実験や調査を行うことは自然科学や社会科学の研究方法の1つの常道であり、心理学も自然科学や社会科学の研究方法を採用していることから、実験や調査は研究の重要な手段となっている。そして、実験や調査によって得られたデータに所定の解析を施すことによって、人間の心的特徴に関する何らかの知見を引き出すのが、科学としての心理学の基本的な研究スタイルである。

そして心理学の場合、実験や調査によって得られる測定値は決まった値になるわけではなく、ある程度のばらつきを持ったものになる。また、それらの測定値はあくまでも標本の値であり、標本を抽出した際の母集団の特徴がそのままデータに反映されているとは限らない。したがって、心理学の実験や調査によって得られたデータにおいては、データに見られる特徴を母集団の特徴として一般化してよいかどうかの確認が必要になってくる。そのための主たる手段の1つが有意差の検定である。有意差の検定方法にはさまざまなものがあるが、それらの狙いとするところは、データに見られる条件間の差が、単に実験対象や調査対象となった人たちだけのものではなく、その人たちを含む何らかの集団の全体に見られる差であることを示すことにある。そして心理学においては、特にことわりがなければ、その集団とは人間全体であるのが普通である。そのような壮大な狙いをもって検定がなされ、検定の結果次第でその狙いが一応は果たされることになるのである。

しかしながら、そのようなことはたやすくなされることなのであろうか。人間全体からすればごく一部のほんのわずかな人間について得られた結果を人間全体に当てはめることなど本当にできるのであろうか。実は、そのようなある意味で無謀とも思えることを遂行するに際して、いくつもの統計法上の仮定が設けられているのである。すなわち、心理学では多くの場合、実験や調査が行われると、その結果について検定を行って、結果の一般化がなされるが、その手続きの中にいくつもの事柄が仮定されたうえで結論が導かれているのである。そのことは、もちろん、実験データや調査データを扱うにあたっては当然了解されているはずのことである。しかしながら、それらの仮定の設定は、普段はとりたてて意識することなく、ごく当たり前のように行われていることが多い。そこで、本稿では、心理学のデータ解析で用いられている有意差の検定における統計法上の仮定を改めて確認

するとともに、そのような仮定を設けることの研究上の意味や留意すべきことについての考察を試みる。

t 検定について

ここでは、有意差の検定の典型的な特徴を有しているとも言える t 検定を取り上げ、 t 検定の適用において仮定されている事柄を明確化していくことにする。

t 検定は、2つの条件下で得られたそれぞれの平均値の間の差の有意性を検定する際に用いられる方法である。すなわち、条件Aと条件Bのもとでそれぞれ得られた測定値の平均値間に何らかの差が見られるとき、その差が偶然による差ではなく、各条件の測定値を標本とする母集団においても差があることによってもたらされた有意な差である、と言えるかどうかを判定する際によく用いられるのが t 検定である。

ここで1つの例として、単語を記銘材料とした記憶の実験の場合を取り上げてみる。実験のやり方はたとえば次のようなものである。まず、たとえば15個の単語を記銘材料として用い、その単語を実験参加者に提示して覚えてもらう。その際に、ある実験参加者にはAという覚え方で覚えてもらい、別の実験参加者にはBという覚え方で覚えてもらう。それぞれの覚え方で覚えてもらいながら、15個の単語を提示し終えたのち、どのような単語があったかを思い出して再生してもらう。そして正しく再生された単語の個数を記録する。このような手続きを、A、Bいずれかの覚え方の条件下でそれぞれ20人程度の実験参加者について行い、各実験参加者が正しく再生した単語の個数の平均値をA、Bの各条件ごとに求める。そして、条件間で平均値の差が見られるならば、その差が有意な差であると言えるかどうかを、 t 検定を行って調べる。そして、 t 検定の結果、有意な差であると判定できれば、覚え方の違いによる効果が認められたものとみなされる。以上は、ごく単純ではあるが、記憶の実験の典型的かつ基本的なケースであると言える。

測定尺度の仮定

まず、 t 検定を行うには、測定値が間隔尺度以上の尺度で測定されたものでなければならない。間隔尺度においては、数値と数値との間の差の単位が数値の位置に関わらず一定でなければならない。すなわち、10と11の間の差も5と6の間の差も同じでなければならないのである。

再生された単語の個数という面では、10個と11個の差の1個も、5個と6個の差の1個も単語1個であることに変わりはない。すなわち単語の個数を物理的な観点のみから見れ

ば、一定性は認められる。しかし、記憶の実験材料として見たとき単語1つ1つの覚えやすさは、必ずしも同じではない。単語の覚えやすさは、熟知度、有意味度、具体性－抽象性、品詞、音節数、表記文字、等々の特性によって異なっている。したがって、記憶の実験材料としての単語は1個1個が必ずしも同等ではないのである。もちろん、それらの違いがなるべく影響しないように、実験に使用する単語の諸特性の値が一定の範囲内に収まるように努めるべきであるし、実際に可能な限りでの統制は行われているが、1つ1つの単語が完全に同等になるように統制を行うことは不可能である。

さらに、そういった材料の問題だけではなく、再生された単語の個数が記憶の心理量を反映するとみなす点においても、測定値の単位が数値の位置に関わらず一定であるとはいえない面がある。たとえば、上述のような実験の場合、0個と1個の間の差と14個と15個の間の差はいずれも1個であるが、果たしてその差に反映された心理量は同等だと言えるであろうか。すなわち、1つさえも思い出せない場合と1つぐらいは思い出せる場合との差と、1つだけどうしても思い出せない場合とすべて思い出せる場合との差が、心理的に同等であるとは考えにくい。常識的に考えれば、覚えた単語のうち1つぐら思い出すことは容易であるが、最後の1つを思い出すのはかなり難易度が高くなるであろう。であるとすれば、0個と1個の間の1個と14個と15個の間の1個は心理量としては必ずしも同じであるとは言えないのである。

以上のようなことを踏まえるならば、単語を記憶させて再生させ、その再生個数で記憶の度合いを測定する、という実験で得られた測定値は、間隔尺度によるものとは言い難い面があるのである。しかし、それにも関わらず、実際には多くの場合、得られた測定値が「間隔尺度による測定値であること」を仮定したうえで、平均値を求め、 t 検定を適用しているのである。

正規分布の仮定

t 検定は t 分布という確率分布を使った検定である。 t 分布のもとになっているのは、正規分布という確率分布である。正規分布する母集団の標本平均の分布も正規分布するが、標本平均の分布の得点を標準化する式における母分散を、その不偏推定値に置き換えた式による得点の分布は t 分布になる。そして、正規分布する2つの母集団の平均値の差が0であり、かつ分散が等しいとき、標本の平均値の差を標準誤差の推定値で割った値が t 分布に従うことを利用して行うのが t 検定である。

したがって、 t 検定を行うときは、対象とする2つの母集団はそれぞれ正規分布するこ

とが仮定されている¹。そして、心理学的な測定値に限らず、一般的に人間の身長や体重、学力、知能といった人間の心身に関する多くの測定値は正規分布に従うことが仮定されている。確かに、実際にそれらの測定値を多数収集すれば正規分布に近似してくることは認められている。しかしながら、正規分布自体はあくまでも、関数式によって定義される理論上の分布である。心理学的な測定値の集まりが正規分布に近似するとは言え、あくまでも近似するだけであり、いくら多数の測定値を集めても正規分布に完全に一致するわけではない。正規分布が理論上の分布であり、現実の測定値の分布が決して正規分布そのものになりえないことは、正規分布する変数 x の値が $-\infty$ から $+\infty$ まで至ることを挙げるまでもなく、明らかなことである。したがって、「心理学的な測定値は正規分布に従う」という仮定を設けたうえで、正規分布から導かれた t 分布を利用しているのである。²

無作為抽出の仮定

t 分布は、母集団の標本平均の分布である。すなわち、母集団から個数を決めて抽出した標本の平均値の分布である。そして t 分布においては、標本の抽出が無作為であることが前提となっている。無作為抽出とは、抽出される個々の標本が、相互に独立に、偶然のみによって決まるようにすることにより、抽出された標本が母集団の中で偏りがなくなるようにすることである。しかしながら、実際の測定値を無作為抽出によって得ることは現実には無理である。通常は実験参加者は、実験者が依頼できる範囲で集められるのが普通である。実験者が大学に所属している場合は、その大学の学生が実験参加者となる場合が多く、人間全体からすれば、著しい偏りがあることは明らかである。そもそも人間全体を対象にした研究において、実験参加者を無作為に集めることなど無理である。心理学の研究対象は、特にことわりがない限りは人間一般ということになるはずであり、そうであるならば母集団はすべての人間についての測定値ということになるが、その中から無作為に標本を得ることは不可能である。なるべく偏りがないように努力や工夫をすることは大事であるが、そのための労力に見合った成果が得られるわけではない。むしろ、実験参加者があまりに多様であると、測定値のばらつきが大きくなりすぎて実験条件による効果が検出されにくくなることさえ生じるのである。したがって、そのような無作為抽出に近づけるための努力や工夫は可能な範囲にとどめ、「無作為抽出であること」を仮定して t 分布を使用しているのが実情なのである。

¹ 2つの母集団の分散が等しいことも仮定されている。

² ただ、母集団が正規分布しなければ t 検定の使用が認められない、というわけではない。母集団が正規分布しなくても標本の個数が大きくなるにつれて、標本平均の分布が正規分布に近づいていくことが知られている（中心極限定理）ことから、たとえ母集団が正規分布でなくても t 検定の適用はある程度可能である。

帰無仮説の仮定

t 検定を行う場合の帰無仮説は「母集団の平均値には差がない」というものである。そしてその帰無仮説のもとで、実験で得られた測定値の平均値間の差と標準誤差の推定値とから、 t の値を算出し、その値が t 分布の中でどのくらいの確率で生じうるのかを調べる。そして、その確率が極めて小さいものであることが確認できたら、帰無仮説を棄却し、「母集団の平均値には差がある」という対立仮説を採択する。そして、測定値の平均値間には有意差があるとし、平均値における差は偶然に生じた差ではなく母集団でも差があることを反映した差である、と判定するのである。よって帰無仮説は、「母集団の平均値には差がある」という対立仮説の妥当性を立証するための仕掛けのようなものである。母集団の平均値には差がある、と言うための根拠を合理的に示すために、母集団の平均値には差がない、という仮説が成り立つ可能性の低さを示し、それに基づいて帰無仮説を棄却することによって母集団の平均値には差があることの根拠とするのである。

ここで、帰無仮説における「母集団の平均値には差がない」ということは、2つの母集団の平均値が全く同一である、ということである。しかし、実際にはそのようなことはまず起こりえない。したがって、実際には起こりえない場合を起こりうるものと仮定し、その仮定のもとでの測定値の平均値間の差の生じにくさを示しているのである。すなわち、ここに「母集団の平均値には差がない場合がありうる」という仮定が設けられているのである。

有意差検定の意義

以上のように、 t 検定による有意差の検定には、測定値の尺度に関する仮定、母集団に関する仮定、標本抽出に関する仮定、帰無仮説に関する仮定、などいくつもの仮定が含まれている。検定の結果はいくつもの仮定を設けたうえで引き出された結論であり、仮定することが正しいという前提のもとで有効な知見なのである。その点から言えば、データに見られる差やそれに対する検定結果はあくまでも暫定的なものと言わざるを得ない。

有意差の有無を検定することの意義は、データに見られる条件間の差を不用意に一般化することを避けることであろう。母集団では全く差がない、という帰無仮説のもとでもある程度の確率で生じうる差であれば、そのような差は偶然に生じうる差である可能性を排除できない、ということである。ある程度の確率の下限を、心理学で通常採用される基準である5%にするのが適当かどうかは議論の余地があるが、偶然に生じた差である可能性をなるべく厳格に排除することは必要であろう。数値上は差が見られていても、それはあ

くまでも標本に見られる差であって、それだけを根拠に母集団のことに一般化して論じるのは適当ではない。その標本に偶然に生じた差であれば、母集団に当てはめることはできない。そこで検定を行うことにより、偶然に生じた可能性を排除できない結果については、母集団への一般化を見合わせることにするわけである。

理論の利用

正規分布も t 分布も、関数式によって定義される理論上の分布である。そして、検定においては、実験や調査によって得られた実際の測定値の集まりに、理論上の分布を当てはめることによって、実際の測定値に表れていることの意味を推測しようとするのである。

このようなことは、実は、研究活動の多くに共通する営みである。もちろん、現実の有様を観察することは研究において重要な営みである。そして、現実の有様をよく観察することによって、表面に表れていないことについてある程度推測することも可能であろう。しかしながら、その推測は長年の経験や直観だけを頼りに行うのではなく、理論を当てはめることによって行うこともできる。すなわち得られた観察結果に何らかの理論を当てはめることによって、観察結果だけでは分からないことを、推測することができるのである。もちろん、理論は人間が思考活動によって作り出したものである。したがって、現実と合っていない部分や現実とズレている部分があるのは当然である。理論を当てはめて得られた推測が間違っていることも十分にありうる。しかし、それでもなお、現実には理論を当てはめることによって得られる推測の価値は多分にある。理論と現実異なる、そのことを承知の上で、現実の観察だけでは分からないことを推測するために理論が用いられるのである。

それと同じことが統計的検定でも行われている。正規分布や t 分布といった理論上の分布を、現実の測定値の集まりに当てはめることによって、測定値に表れている事柄の意味を推測するのが、統計的検定の本質的な意義である。 t 検定とは異なり、母集団が正規分布することを前提としないノンパラメトリック検定も数多く考案されている。それらの検定では母集団の正規分布についての仮定は不要になるが、それでもなお、帰無仮説の棄却は、標準正規分布や χ^2 分布、その他何らかの確率分布を用いてなされる。現実の測定値の集まりの意味しているものを、理論を用いて推測するという点は、ノンパラメトリック検定の場合にも共通していると言えるであろう。

尺度の問題

心理学のデータ解析において、1つの大きな課題と思われるのは、測定値の尺度の問題である。 t 検定や分散分析といった心理学でよく用いられる検定法は、測定値が間隔尺度以上の尺度によるものであることが条件である。その条件を満たすため、心理学の実験や調査ではさまざまな工夫や努力がなされている。人間の行動や内面を数量的に測定するための工夫や努力を最も熱心に行っている学問はおそらく心理学であると言ってもよいかもしれない。しかし、たとえそうであっても、心理学の実験や調査で得られた測定値は間隔尺度としての条件を十分に満たしているとは言い難いのも事実である。そのため、 t 検定や分散分析を行う際には、間隔尺度による測定値であることをあくまでも仮定して行っているのである。

その点に関し、ノンパラメトリック検定の多くは、順序尺度や名義尺度による測定値に適用できる検定である。しかし、ノンパラメトリック検定は概して有意差の検出力が低く、また、順序尺度や名義尺度による測定値は、間隔尺度による測定値に比べてデータの示す内容の情報量が少なくなる。そのようなことを嫌って、多くの場合、間隔尺度としての条件を必ずしも満たしていないにも関わらず、間隔尺度によるデータであることを仮定して検定を行っているのである。このことは十分に承知しておくべきことであろう。

検定を行う以前に、平均値や標準偏差などの基本的な統計量の算出さえも、本来、間隔尺度以上の尺度による測定値でなければ認められない。そして、心理学の実験や調査で得られた測定値の大部分がその条件を必ずしも満たしていない、とするならば、人間の行動や内面を数値で表すことや、それに基づいて何らかの合理的な結論を引き出すことが、いかに難しいことかを十分にわきまえることも重要であろう。安易なやり方で数値化して、その数値を一人歩きさせたり、その数値だけで簡単に人間の行動や性質を評価したりすることは、可能な限り慎むべきではないであろうか。

参考文献

- 森 敏昭・吉田寿夫(編著)(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
- モリソン, D.E.・ヘンケル, R.E.(編)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
- 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

吉田寿夫(1998). 本当にわかりやすすぎく大切なことが書いてあるごく初歩の統計の本
北大路書房