

教育経営研究における統計的手法の留意点 (1)

—検定における効果量とサンプルサイズについて—

植 田 義 幸

教育経営の分野では近年、教育的働きかけや教育施策、行政的条件の結果を数量的に扱い、多様な統計的手法によって分析が行われるようになってきている。しかし、高度な手法が用いられていても、その結果に対する解釈は理解困難なものもある。このような統計手法の基礎となっている統計的検定について、効果量とサンプルサイズの二つの側面から、教育経営学での適切な扱い方を検討した。効果量は統計的有意性ととも報告することが求められたり、多くの研究の成果を統合するためのメタ分析で用いられったりする概念でもあり、今後は研究結果報告の中で記述することが望ましい。分析手法によって効果量の算出法は異なるが、Cohenによる効果量の目安を紹介し、研究成果の有効性の判断基準として活用することを提案した。サンプルサイズについては、有意水準や検定力、効果量の大きさから、適切な大きさが算出できる。小さすぎるサンプルは研究者の害になり、大きすぎるサンプルは検定の意味をなくしたり、被験者に無用な負担を強いたりすることになるため、適切なサンプルサイズを意識して研究を行うことを提案した。

キーワード：検定、効果量、サンプルサイズ、全数調査

1. はじめに

近年の教育経営学における研究では、多様な統計的手法が用いられている。露口 (2007) によれば、同稿で調査対象とした53編の研究論文のうち、カイ二乗検定 (22.6%)、t 検定 (22.6%)、分散分析 (24.5%)、相関分析 (28.3%)、回帰分析 (30.2%)、共分散構造分析 (11.3%) が用いられており、「解析技法のタイプとしては、アメリカ等の諸外国における教育経営研究とほぼ同様」という状況である。また、初期にはカイ二乗検定やt検定が多かったのに対し、近年は回帰分析、共分散構造分析、ロジスティック回帰分析、順序回帰分析等が活用されているとする。教育経営研究においては、経済学における回帰分析、心理学における因子分析や分散分析のような、支配的な手法は見られないことがわかる。

この多様さは、対象や関心の多様さに応じるものであろう。様々な教育的働きかけや教育施策、行政的な条件の影響について、国同士、都道府県や市町村間の比較、学校を対象とする研究、教員、児童生徒や保護者といった個人を対象とする研究などがあり、関心の対象も、財政状況や人口、各種の指標、学校の組織状況、個人の心理的な側面、能力や学歴等にわたり、正規分布等の既知の分布でほぼ近似できるものもあれば、どのような分布を仮定してよいのかさ

え不明なものも多い。

教育経営研究における統計的分析を理解するには、その手法について高度な知識が求められている。さもないければ、解釈が正しいのかどうかさえ判断できない場合もある。このような、多様で複雑な統計的手法を理解するために、これらの分析相互の関連を考察する必要がある。多様な手法を比較する視点が必要になっている。

手法は様々だが、その大多数において、対象を調査したり、既存の集計を利用したりして、単に集計するのみならず、その数量についていくつかの群間の差や変数間の関連が偶然ではないことを示すために統計的検定（以下では「検定」とする）が行われている。本稿では、複雑な手法の基礎となっており、多くの研究で行われている検定のあり方を検討してみたい。ただし、本稿ではサンプルに対する研究（調査や実験）の結果得られたデータを用い、帰無仮説（たとえば「サンプルの平均は母集団の平均と差がない」という仮説）を一定の基準で棄却するか採択するかを判断するタイプの検定を扱うことにする。検定には、得られたデータの分布が既知の分布に適合しているかどうかをテストする適合度検定等もあるが、これらは「仮説が棄却されない」ことが望ましいタイプの検定であり、多少異なる視点からの議論が必要であるため、本稿では割愛する。

検定では、サンプルをどのようにして手に入れるか、あるいはどのようにして手に入れたサンプルであるか、有意水準は適切かといった問題が取り上げられ検討されることはある。しかし、実際の検定手続きの中では、どの程度の差があるかという点が充分検討されず、「有意かそうでないか」、「仮説が棄却されるか採択されるか」が過大に重視されているように思われる。さらに、「有意である（差がないという帰無仮説が棄却された）から差があると言える」「差があるから群分けや独立変数の効果がある」というロジックが見られる。

ここでは、有意になるかどうかを左右する「効果量（effect size）」（効果サイズやエフェクトサイズと表現されることがあるが「効果量」で統一する）とサンプルサイズ（標本数や標本サイズと表現されることがあるが「サンプルサイズ」で統一する）について述べたい。もちろんサンプルをどう得るか、変数はどのような分布か、有意水準を何%にするか、といった点も重大な問題であるが、本稿では、効果量や適切なサンプルサイズの考え方の基礎を紹介し、その意味するところを検討し、教育経営研究の推進に資することを目指したい。

2. 教育経営研究における検定の妥当性

2. 1. 効果量

冒頭で述べたように、教育経営研究における統計手法として、変数間の構造を探る手法として、因子分析や主成分分析、コレスポンデンス分析（数量化Ⅲ類）なども多く用いられているが、t検定のような2群の差の検定や分散分析、回帰分析、その発展形である重回帰分析、ロジスティック回帰分析、順序回帰分析、生存分析といった高度な手法も用いられている。これらは予測するための変数（独立変数、予測変数、説明変数）が、予測される変数（従属変数、目的変数、基準変数）に与える因果的影響を検討するものである。これらの手法に共通した判定基準として、本稿では「効果量」が十分な意義があることを示す。

次のような架空の事例を考えてみる。教育経営学に関する授業の受講学生に対し、2変数からなる400件のデータを与えた。データは、それぞれ200人からなる B_0 、 B_1 の2群の被験者に対し、ある設問への回答を7件法のリッカート尺度(A尺度)で評定することを求めたものとする。群(ダミー変数)と評定との関連について2変数間の関連を調べよという課題を出したところ、表1～表4の4種類の表を中心としたレポートが提出されたとする。それぞれ、別途基礎統計量などは報告されているものとする。

表1 相関分析の例

与えられたデータ(各群200人、全体平均=3.95)を用い、群をダミー変数としたカテゴリーとA尺度との相関係数(点双列相関係数)を求めたところ、 $r=0.124^*$ であった。有意水準5%で統計的に有意な相関があることがわかった。

表2 t検定の例

	全体 (n=400)			第1群 (n=200)			第2群 (n=200)			t 値
	平均	SD	SE	平均	SD	SE	平均	SD	SE	
A 尺度	3.95	1.03	0.052	3.82	1.02	0.072	4.08	1.03	0.073	-2.48*

表3 分散分析の例

要因	自由度	平方和	平均平方	F 値
群	1	6.50	6.50	6.17*
誤差	398	419.39	1.05	
合計	399	425.90		

表4 回帰分析の例

予測変数	係数	標準誤差	β	t 値
定数	3.82	0.073	—	52.63*
群	0.26	0.10	0.124	2.48*

回帰式： $y=3.82+0.26x$
決定係数 R^2 : 1.5%

表1は質的変数である群と量的変数とみなしたA尺度との相関係数(点双列相関係数)を調べたものである。ここでは表形式にする必要はないため、文章形式で掲げている。表2は、2群の差の検定(t検定)によって群間の差を検討するものである。ただし、両側検定によるものとする。表3は、表2と同じく群間の差を検討するが、一元配置の分散分析(ANOVA)によって分析した結果である。表4は、群をダミー変数化(B_0 群を0、 B_1 群を1とする)し、A尺度を従属変数にして回帰分析した結果である。

ただし、いずれの表でも有意水準を5%とし、有意な結果に「*」を付している。すべての

手法において有意水準 5 % で有意だという結果が報告されていることになる。

どれを最も適切な分析手法であるとして評価すべきであろうか。t 検定は 2 群の差の検定にしか用いることができないから、表 3 の分散分析の方が応用範囲が広く、その後の学習につながるので適切であるといった「教育的見地」からの評価ではなく、統計的手法として正しいのはどれだろうか。

この 4 つの分析は本質的には同じことを表している。表 1 に示す相関分析に関しては次のような見方があるだろう。「相関係数が 0 であるという帰無仮説は危険率 5 % で棄却されるが、その相関係数は 0.124 であり、非常に弱い関連しか認められない。これを相関ありと報告するのは問題がある」と評価すべきだと判断する人がいるかもしれない。初学者向けの入門書では相関係数の絶対値が 0.2 以下なら「相関なし」、0.2 ～ 0.4 で「弱い相関あり」、0.4 ～ 0.7 で「相関あり」、0.7 以上で「強い相関あり」とであると述べるものもある（内田・平野 (2011;p.85)）。森・吉田 (1990;pp.230-231) によれば「無相関検定は、あくまで 2 変数間に直線関係が存在するか否か、すなわち $H_0: \rho=0$ に関する検定である。したがって、検定の結果がたとえ高い水準で有意（たとえば $p<0.001$ ）であっても、それは必ずしも両者の間に強い直線関係が存在することを意味しない」、「2 変数間にどの程度強い相関関係があるかは r の値そのものが表すのであり、単に無相関検定だけを行うのではなく」「 r の信頼限界を求めておくことなどが望まれる。さらに、単に r が有意であることから、その関係を絶対視し、まるで $|r|=1$ であるかのように結果を記述することは、当然慎まなければならない」としている。これは正しい扱いである。しかし、それは表 2 ～ 表 4 に示した他の分析でも同様である。表中には記さなかったが、有意確率は表 1 ～ 表 4 のすべてで $p=0.013$ で等しい。

たとえば、表 4 の回帰分析は 2 変数に関連があることを示しているように見えるが、実際には、報告された内容は、評定値と群との間には「ほとんど関連がない」ことを示しているとも言える。すなわち、決定係数は 1.5% ($=0.124^2$) でしかない。表 1 の相関係数に対する判断を「相関が非常に弱い」もしくは「相関はない」とするのであれば、表 4 に示す回帰分析に対しても同じ判断をすべきである。他の t 検定と分散分析についても同様である。表 1 の相関分析に問題があると評価するのならば、他の分析についても同じ判断をして、「弱い関連しかない」結果を意味ありげに報告するのは間違っていると評すべきである。

表 2 ～ 表 4 の分析（t 検定、分散分析、回帰分析）は表 1 の分析（相関分析）よりも多様な情報を含んでいる。t 検定では、各群の平均や標準偏差がわかるし、分散分析ならば平方和や平均平方などが掲げられている。その点を評価すべきであるという見方もあるが、それらは基礎統計量として別に報告すればよいのであるから、これらは本質的に同等のものであるという判断が求められる。

このような、本質的に同等であるにも関わらず表現が異なる手法同士を比較したり、異なる条件のもとで行われた研究同士を比較・統合したりするために「効果量」という指標をあわせて検討する必要があるという主張がなされるようになってきている。研究報告に効果量の記載を求める動きもある。APA（アメリカ心理学会）の論文作成マニュアル第 6 版（原著 APA Publication Manual (6th ed)）では「研究で得られた知見の重要性を読者が完全に理解するには、

「結果」のセクションに効果量という尺度を加える必要がある」と述べる。

効果量は、2 群の差の大きさや変数間の関連の強さそのものを表している。検定はサンプルの統計量によって母集団の様々な統計量（母数）がどのようなものであるかを推測して判断する手法であり、サンプルが母集団を適切に代表しているかどうかを判断する手法であると言い換えられる。それに対し、効果量は、推定される母数によって、問題となっている変数の群間での差や変数間の関連の度合い、あるいは独立変数が従属変数に対してどの程度の影響を持つかを検討するためのものである。

先に取り上げた分析方法と、その方法で用いられる効果量に対する呼称、効果量の大きさがどの程度の効果だと判断すべきかの目安を表 5 に掲げる。Cohen (1988) が示した効果量の大きさの目安を Ellis (2010) がまとめたものから抜粋して引用した。

表 1～表 4 で示した各手法について、効果量で表現すれば、相関分析 (表 1) は 0.124、t 検定 (表 2) は 0.256、分散分析 (表 3) と回帰分析 (表 4) は 0.015 である。いずれも表 5 に照らすと、せいぜい「小さな効果」を少し超える程度である。

表 5 統計手法と効果量の目安

手法	呼称	効果量の目安		
		小	中	大
相関分析	R	.10	.30	.50
t 検定	D	.20	.50	.80
分散分析 (1 元配置)	η^2	.01	.06	.14
回帰分析	r^2	.02	.13	.26

Cohen (1988) に基づき Ellis (2010) が作成した表より抜粋

表 5 では t 検定の場合に d (Cohen's d) が掲げられているのみであるが、他にもいくつかの差の大きさを表す指標がある。これらは、差を標準偏差で割ったものであるため、特定の範囲にはとどまらないが、おおむね 0～3 程度におさまる。分母の標準偏差として何を採用するかによって Glass's delta、Hedges' d 等の複数の指標があり、Cohen (1988) では「d ファミリー」と称されている。

一方、相関分析、回帰分析、分散分析で用いられる指標は「r ファミリー」と称される。これらの関連の強さを表す指標は、回帰分析における決定係数のように 0 から 1 の値をとる。「d ファミリー」に属する指標も「r ファミリー」に属する指標も、有意確率のようにサンプルサイズが大きくなると小さくなるような性質は持たない。

Mullen (2000 (原著 1989) pp.50-55) は「研究結果のひとつの次元は有意水準である」とし、「もうひとつの次元は効果サイズ (effect size)」であるとする。そして「ときどき学者は有意水準あるいは効果サイズのいずれが有益か、あるいは適切かという議論を戦わせたがる」が、「この議論 (中略) は有意水準と効果サイズのエレガントに単純な相互依存関係を理解していない」と批判する。「どんな有意性検定も効果サイズと研究サイズの関数として表現できる」として「有意水準 = 効果サイズ × 研究サイズ」と模式的に示している。また、南風原 (2002;p.163) では「検

定統計量＝効果の大きさ×標本の大きさという関係が成り立っていることがわかります」としている。つまり、サンプルサイズが大きくなれば、効果量が一定でも有意になりやすく、サンプルサイズが一定なら効果量が大きくなれば有意になりやすいことが示されている。

教育経営分野においても、研究者は、帰無仮説が棄却されたことを示すだけでなく、先に示した目安等にしたがって、その研究が示す結果がどの程度の効果を示すのかを研究報告に含めることが望ましいのではないだろうか。

また、効果量は、最近ではメタ分析との関連で取り上げられることが多い。近年、医学や薬学の分野では、同じ疾病等に対して同じ治療法や薬剤を用いた独立した複数の研究を統合し、全体としてその治療法や薬剤の有効性や毒性を評価するためにメタ分析という手法が広まりつつある。同じ薬剤を用いて同じ疾病に対して行われた治験などのサンプルサイズが異なるような場合でも比較可能になるように、あるいは母集団が多様な複数の研究を統合する際の基準として、効果量が用いられている。

今のところ、教育経営研究では同一の関心に基づき、同一の教育手法や施策を異なる多数の研究グループが別々に取り上げて研究するという状況はまれなことであろう。しかし、類似した状況間の比較が必要となることはあり得るし、ある研究者・研究グループが同種の研究を経時的に実施したときに、サンプルサイズの規模によって、あるときは有意となり、あるときは有意とならないこともある。このようなときにはサンプルの質の違いを検討するだけでなく、複数の結果を比較検討する基準が必要となるため、効果量を示すことも有効となるだろう。

2. 2. 検定におけるサンプルサイズ

次に、サンプルサイズの問題について検討してみる。冒頭で述べたように、多くの研究で検定が行われるが、「有意な」結果を得るには、効果量が大きい調査や実験を行うことを優先すべきである。たとえば、異なる教育方法を施した2群の生徒の成績が有意に異なるかどうかを確かめるためには、従来の方法よりも大きな有効性をもつ新しい教育方法を用いて調べれば、2群の成績に大きな差がつき、検定でも「有意」になるはずである。しかし、有意な差が見られなかった場合に、しばしば「サンプルが足りなかった」との意見（ないしは反省）を聞く。「サンプルが足りなかった」ために測定誤差が大きくて有意になるはずの差を確かめられなかったとの趣旨であれば了解できるが、「サンプルが足りなかった」ために自由度が小さかったためとか、標準誤差を小さくできなかったためとかいった理由で検定量が有意になるべき数値に達しなかったという意味であれば本末転倒ではないだろうか。初学者に対し、「サンプルを増やせば有意にできる」といったアドバイスをするのは適切ではない。

検定はサンプルによるデータから母数を推定する考え方の応用である。推定の誤差を表す標準誤差（SE）はサンプルサイズが大きくなればその平方根に反比例して小さくなる。4倍の大きさのサンプルを用いれば、SEは2分の1になる。たとえば2群の差の検定であれば、およそSEの2倍の差があれば「有意差」があると判断される。したがって、サンプルサイズが大きくなれば有意な結果が出やすくなる。検定の種類によっては適合度の検定のように有意と

ならない方が研究者にとっては望ましい場合もあるが、多くの場合は差はないとする帰無仮説が棄却されることをめざすため、研究者はサンプルサイズをなるべく大きくしようとする傾向がある。10人から得たデータでは有意とならないが、同じ差でも、100人、1000人と被験者を増やせば、帰無仮説が棄却される確率は高まる。たとえば7件法のリッカート尺度で得た二つの群（標準偏差1程度）の平均の差が0.2しかなくとも1000人の被験者からデータを得ればおそらく「有意差がある」ということになるだろう。それは統計学的には「誤り」ではないが、7件法での平均が、ある群では3.9、別の群では4.1であるという事実には意味があるだろうか。

そのためには、固有科学（あるいは固有技術）的に、つまり実質的に「意味がある」か「意味がない」かを先に検討し、「意味がある」差が認められた場合には統計的に差があるかどうかの検討を開始するという手続きが必要であろう。実質的に意味がない差を、統計的検定によって「有意差がある」と述べることは無意味であり、有害でさえある。

サンプルサイズについて、永田（1996;pp.113-114）は「ヒストグラムを描けるほどデータ数があるのに検定する」のは誤用であると述べている。具体的には「 n_1 と n_2 のそれぞれが約40～50以上の場合」には「形式的に検定をおこなって実務的に意味のない差を有意差ありとみなすよりも、点推定値を比較して考察するほうがよい」、「検定は、小試料（サンプルサイズが小さい場合のデータ）に対しておこなう統計的推測の手段である」と述べる。また、服部・海保（1996; p.97）は「統計的検定について一般にいえることですが（中略）標本を増やしていけば必ず帰無仮説は棄却されてしまいます。（中略）意味のある差がなくても、標本の大きさが大きくなれば、やはり帰無仮説は棄却されてしまいます。つまり、統計的な有意差が、実質的に意味のある差があることを保証するわけではありません」と述べている。したがって、サンプルサイズが小さく、差の有無が不確実な場合は、有意性検定によって有意な差があるのかないのかを検討すべきであるが、サンプルサイズが巨大なものである場合は、検定力が高くなりすぎるため、統計的有意差があることが実質的に意味があることを追認するに過ぎなくなることがある。また、上に述べたように、実質的に意味のない差を「統計的に」有意であることが示される場合もある。いずれの場合も検定をすることは不適切であると考えられる。

例えば、ある工場で作る製品は平均10000時間（SD=200時間）の寿命を持つものとする。最近の検査で、サンプル200個を検査したところ、（何らかの原因で）寿命が50時間延びていることがわかった。この時、有意性検定を実施すると（ $z=3.54$ 、 $p=.000$ ）となり、平均寿命は有意に長くなっていることがわかる。しかし、10000時間もの寿命がある部品の寿命が50時間延びることに実質的な意味があるかどうかは、この製品の用いられる状況、原材料や加工のコストなどを総合的に勘案しなければわからない。一方、寿命が100時間（SD=2時間）の製品であれば、50時間の寿命延長は画期的であろう。検査を行い、検定を実施する意義はある。しかし、前者のような状況では、検査は必要であるとしても、わざわざ検定する意義があるのかどうかは固有技術に基づいて検討しなければならない。

適切なサンプルサイズを計算する方法を示す。一般に検定では「実際には差がないのに差がある」と誤って判断してしまう（第1種の誤り）確率を α 「実際には差があるのに差がない」と誤って判断してしまう（第2種の誤り）確率 β を考える。多くの分野では α を0.05（5%）

とする¹⁾。 β については言及しないことが多いが、 $\beta=0.2$ として、検定力 ($1-\beta$) を0.8とすることが多いようである。 $\alpha=0.05$ と $\beta=0.2$ を前提にして考えてみることにする。

サンプルサイズは、有意水準 (α)、検定力 ($1-\beta$)、効果量を定めれば求めることができる。たとえば、相関係数について、有意水準を0.05、検定力を0.8、効果量 (r) を0.3として中位の効果量とする。SYSTAT13(SYSTAT社)を用いて求めると、標本数は85必要となることがわかる。85件のサンプルを集めるには、児童生徒学生や保護者を対象とする調査研究ならば可能である。児童生徒学生や保護者を対象にした調査を行うと、300程度以上のサンプルが容易に得られることも多い。少し大規模な調査を行えば、数千ものサンプルが得られ、大きすぎるサンプルとなる。

しかし、5件法や7件法で調査をしたとき、100人のうち一人が1単位高い答えをすると、平均値は0.01上昇する。たとえば、9割の人は同等で、1割が2→4と答えるようになれば、平均値は0.2異なる。また、一人がスケールを誤って1と答えるべき回答を逆に5と回答すると、平均値は0.04上昇する。どの程度異なると「意識が異なる」と認識できるかは明らかではなく、固有科学的な検討が必要である。

サンプルサイズについては、市町村やその教育委員会を対象とした調査研究等でも同様である。しかし、都道府県やその教育委員会を対象とした研究では、この効果量と有意水準では仮説を棄却することは不可能だということになる。サンプルサイズの上限が47に制限されるため、有意水準0.05、検定力0.8の下では、効果量は0.4が必要である。0.4未満の関連の強さでは、都道府県を対象とした研究では適切な検定をするには不適當であることになる。

さらに、先に述べたように、検定とは「母集団」からとられたサンプルによって母数を推定する手続きの応用である。したがって、全ての都道府県を調べてしまえば検定という考え方そのものが成り立たない。つまり、都道府県程度の数しかない対象を扱うときは、実質的な差があるとわかっていても、検定力が小さいため、検定を形式的に実施しても有意差が見いだせない可能性が高い。また、そもそも全数(悉皆)調査を行ってしまえば、検定の意味が不分明になってしまうため、検定は行うべきではないとも言える。

適切な大きさのサンプルを得られない場合、どのようなデメリットがあるだろうか。充分な大きさのサンプルを得られない場合、つまり少なすぎる場合、その害は主に研究者に及ぶ。研究者は適切なサイズなら得られたはずの有意な結論を得られない。つまり、本当は差や関連があるのに、その差や関連を「統計的に有意であった」と結論づけることができないというデメリットを受ける。しかし大きすぎるサンプルの場合、研究者は容易に「有意」な結論を得られるというメリットを得るのに対し、被験者や研究対象は無用なデータを提供させられることになる。医学分野や薬学分野であれば、新薬の治験のように(同意があるとは言え)本来であ

1) 0.01 (1%) や、より厳しく0.001 (0.1%) とする場合もある。また、0.1 (10%) として緩い基準にすることもある。ただし、p値が0.05以上0.10未満となったときに「有意傾向」なる語が用いられることがあるが、これは不適切な語である。検定はあらかじめ有意水準を定め、その水準に従って、仮説の採用と棄却を決定する手法なので「惜しくも」有意にならなかったという表現も誤りである。

ば得られたはずの適切な治療の機会をのがしたり、二重盲検法の場合のように、くじ引きによって無効な偽薬（プラセボ）を与えられて治療の機会を失ったりするという場合さえある。このような治験は可能な限り少人数で行うべきだと考えられる。教育経営研究で生命や健康に害があるような調査研究等は少ないかもしれないが、アンケートに答える時間や労力を求められるのであるのだから、事前の効果量の推定とそれに基づくサンプルサイズの設定は熟慮すべき問題となり得るだろう。

3. おわりに

本稿では検定を教育経営研究において用いる際の問題点を、効果量とサンプルサイズの2つの視点から検討した。

効果量を検討したのは、統計的な結論と固有科学的な結論との対応を統合的に考えるためである。関心も対象もさまざまな教育経営分野では、異なる研究グループ・研究者の独立した研究を統合するためのメタ分析が通常の研究方法として用いられることはまれであろうが、類似した研究や一つの研究グループや一人の研究者が行った、対象の異なる研究をメタ分析することはあり得る。また、教育経営学の固有科学としての「有意性」が、統計的有意性の有無とは独立に存在することを意識した研究態度が必要であることを理解すべきだと考えた。

サンプルサイズの検討は、最近の研究状況と無関係ではない。個人情報保護の観点から調査研究が困難になっている面もあるが、一方ではICT技術の発達により、より容易に多数のサンプルを集めることができるようになっている。「統計的有意差」や「統計的に有意な効果」を否定するものではないが、サンプルサイズが大きくなると、とりわけ数千から数万もの標本を扱う場合には、実質的に無意味な差が統計的に「有意」となることに留意しなければならない。また行政統計が公開されるようになってきており、都道府県や市町村といった単位では対象のすべてについてのデータが得られることもある。全数調査の結果を形式的に検定することは無意味であることも心得ておきたい。

本稿では取り上げることができなかったが、不適切な手法を用いた例は他にも考えられ得る。たとえば、分割表に対してカイ二乗検定を行っているのに、実数ではなく割合（％）のみを示した表のみを掲げたり、（多重比較法を用いずに）多くの項目、たとえば、数十項目の類似した設問への回答状況について検定し、そのうちの何項目かに有意差があるとしたり、高次項を含めて重回帰分析を行い、最高次の項の係数だけでなく、低次の項の係数をも解釈したり、打ち切りデータがない上に全数調査であるのに生存分析をして検定を行ったりするのは適切ではない。一方、教育経営学において、統計的手法が多用されるようになった時期には、複数の群（分散分析における水準）に対してt検定やその変法であるWelch検定等の2群間の差を個別に検定する手法を複数回繰り返すことは不適切であるという知識が広く普及していたためか、分散分析を用いるべきところにt検定を複数回用いるといった誤用は少ないように思われる。これらについては、稿を改めて述べたいと思う。

引用文献

- アメリカ心理学会 (前田樹海・江藤裕之・田中建彦訳) 2011『APA論文作成マニュアル (第2版)』医学書院 (原著American Psychological Association 2009 Publication manual of the American Psychological Association (6th ed.) American Psychological Association)
- Cohen, J. 1988 Statistical power analysis for the behavioral sciences (2nd ed.) Psychology Press
- Ellis, Paul D. 2010 "The Essential Guide to Effect Sizes" Cambridge University Press
- 南風原朝和 2002『心理統計学の基礎』有斐閣
- 服部環・海保博之 1996『Q&A心理データ解析』福村出版
- 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
- Mullen,B (小野寺孝義訳) 2000『基礎から学ぶメタ分析』ナカニシヤ出版 (原著Advanced BASIC Meta-Analysis 1989)
- 永田靖 1996「検定の誤用パターン」『統計的方法のしくみ』日科技連 pp.113-116
- 露口健司 2007「教育経営研究におけるサーベイリサーチの動向と課題」『日本教育経営学会紀要』49 pp.202-213
- 内田治・平野綾子 2011『JMPによるデータ分析』東京図書

参考文献

- 足立堅一 1998『らくらく生物統計学』中山書店
- 水本篤・竹内理 2008「研究論文における効果量の報告のために－基礎的概念と注意点－」『英語教育研究』31 pp.57-66
- 村井潤一郎 2006「サンプルサイズに関する一考察」吉田寿夫編著『心理学研究法の新しいかたち』誠信書房 pp.114-141
- 永田靖 2003『サンプルサイズの決め方』朝倉書店
- 豊田秀樹編著 2009『検定力分析入門：Rで学ぶ最新データ解析』東京図書