上越教育大学研究紀要　第31巻　平成24年２月
Bull. Joetsu Univ. Educ., Vol. 31, Feb. 2012

悉皆調査における統計的推測

奥　村　太　一*

（平成23年９月27日受付；平成23年11月２日受理）

要　　旨

悉皆調査（全数調査）において得られたデータをもとに，母数に関する統計的推測が行われることがしばしばある。一般に，悉皆調査ではサンプリングに伴ってデータが確率的に変動することはない。このような状況で統計的推測を行うことに意味があるのだろうか。本研究では，まず悉皆調査において統計的推測を行うことの意義を測定の信頼性を中心に考察した。次に，例としてそうしたデータに対応のある２群の平均値差の検定を適用した場合の問題点をシミュレーションによって指摘したうえで，測定の信頼性を考慮した改善策を提案した。最後に，測定の信頼性が高まるにつれ悉皆調査において統計的推測を行う意義自体が存在しなくなることを改めて指摘し，統計的推測の手法を工夫することよりも測定の質を高く保つことが重要であることを強調した。

KEY WORDS
census 悉皆調査（全数調査），statistical inference 統計的推測，reliability 信頼性，*t*-test *t* 検定

1　悉皆調査と統計的推測

平成19年６月に行われた学校教育法の改正に伴い，各学校において学校評価を行い学校経営の改善と教育水準の向上に努めること，またそうした評価情報を積極的に公開することが規定された。教育の質保証と情報公開の必要性が叫ばれる中で，テストやアンケートを用いた調査結果から子どもたちの学業や学校生活に対する意見などの現状を把握しようとする動きが活発になっている。こうした調査は，一般的に悉皆調査（全数調査）と考えてよい。一般に，研究目的でデータを収集しようとする場合は，そこで得られる知見を一般化したい対象（母集団）が非常に広く定義されるため，悉皆調査を行うのは現実的に不可能である。実際，極めて大規模に実施されていることで有名な国際学力調査であるPISAやTIMSSでさえも悉皆調査ではなく，抽出調査である。わが国において2007年度から実施されている全国学力・学習状況調査も，当初悉皆調査として行われていたものが開始早々抽出調査に切り替えられた。これは悉皆調査が極めて多くの金銭的，人的，時間的コストを必要とするのに対し，多くの場合そこで目的とされていることは抽出調査で十分果たすことが可能だからである。こうした費用対効果に関する事情も相まって，実際に調査と銘打って実施されるものはそのほとんどが抽出調査であると言ってよい。すなわち，母集団から実際にデータ収集の対象者として協力してくれる人々をサンプルとして抽出し，得られたデータの記述結果をもとに母集団に関して何らかの推測を行うというのが一般的な調査の流れである。

一方，学校評価とはある学校において特定の期間に実施された取り組みを評価するものであるから，一般的には実施困難とされている悉皆調査がいとも簡単に行えてしまう。母集団が地域住民であるといった場合は別として，全校生徒が母集団であるとするならばその全員についてデータを得ることは容易である。これは学校に所属する児童・生徒が学校の管理下にあり名簿が整備されていることと，母集団の規模がそう大きくないためである。むしろ，特定の生徒を抽出してテストやアンケートを実施する方が手間もかかるであろうし，第一不自然であろう。これは，近年多くの大学において実施されている授業評価アンケートや卒業生の進路調査などについてもしばしば当てはまることである。このような悉皆調査によって得られたデータについては統計的推測を行う必要はなく，記述をすることでその目的は満たすことができる。データの記述については平均や比率，相関係数といったよく知られている統計的指標を用いることはもちろん，多くの多変量解析がデータの情報を要約するための手法として提案されている（渡部，1988）。また，効果量（effect size）と呼ばれる一連の標準化された記述指標を用いることも結果を有意味に解釈する上で極めて有用である（Cohen，1988）。

しかし，こうした調査の報告書を読むと，悉皆調査が行われているにもかかわらず母数の値について仮説検定（以下，単に検定と呼ぶ）を始めとした統計的推測が行われているケースをしばしば目にする。言うまでもなく検定は母

*学校教育学系

数に関する仮説の真偽を検証するために行われるものであるから，悉皆調査により母集団についてのデータが得られているのなら行う必要はない。しかし，例えば相関係数の値が$r=.24$などと得られているにも関わらず，「検定の結果５％水準で有意ではなかった」といった記述は検定結果の報告において非常によく目にする。このような記述を見ると，母数値が取るに足らない値であるのか（実質０と見なしてよいか）どうかを確認するために検定が行われているかのように思える。当然のことながら，検定にそのような役割はない。一般に，母数値の解釈は統計的というよりもむしろ実質科学的に行われるべきである。少なくとも，検定は通常データがサンプリングに伴って確率的に変動することを前提としているから，悉皆調査のようにデータが変動しない状況ではその機能を果たすことはできない。

2　測定の信頼性と誤差の確率変動

では，悉皆調査においては検定をはじめとする統計的推測を行うことは一切無意味なのであろうか。確かに，前節で述べたように悉皆調査によって観測されるデータの値が完全に固定されている（確率的に変動しない）と考えるのならばそうであろう。しかし，実際には悉皆調査で得られたデータが全く確率的に変動し得ないものと考えるのは早計である。もちろん，悉皆調査である限り抽出調査のような調査対象者のサンプリングにもとづくデータの確率的変動は存在しない。ここで問題になるのは，測定に伴う誤差にもとづく変動である。一般に，ある尺度を用いて測定を行うとき，得られた観測値には多かれ少なかれ誤差が含まれている。古典的テスト理論と呼ばれる測定理論においては，測定の結果得られる観測値xは，真値tと誤差eの和からなっていると想定する（Gulliksen，1950）。すなわち，

$$x=t+e \tag{1}$$

である。ここで，真値とは同じ対象に同一条件で無限回測定を行った場合の観測値の平均である。測定に伴う誤差を生じさせる原因にはさまざまなタイプが考えられるが（南風原，2002a），ここではそうした様々な原因が組み合わさって誤差が確率的に変動するととりあえず考えておく。ここで，真値と誤差が無相関であると仮定すると，観測値の分散σ_x^2は

$$\sigma_x^2=\sigma_t^2+\sigma_e^2 \tag{2}$$

と真値の分散σ_t^2と誤差の分散σ_e^2とに分割される。このとき，信頼性係数ρが観測値の分散に対する誤差の分散の割合

$$\rho=\frac{\sigma_t^2}{\sigma_t^2+\sigma_e^2} \tag{3}$$

として定義される。

測定の信頼性が低いということは，取りも直さず測定に伴って確率的に変動する成分が観測値に多く含まれるということである。とするならば，信頼性の低さに応じて平均や相関係数の母数値について統計的な推測を行うことができる。すなわち，母相関係数は０であるが確率的に変動する誤差によって例えば$r=.24$のような相関係数の実現値が得られたのではないか，あるいは測定誤差を考慮すると母相関係数の値はいくらと考えられるか，といったことがここでの統計的推測に相当する。本稿では，さまざまな統計量を用いた統計的推測のうち，対応のある２群の平均値差に関するt検定を取り上げてこの問題について考えてみたい。対応のある２群のt検定は，教育実践的な働きかけの効果を検証するために事前事後デザインのデータの分析方法として広く用いられている。

3　信頼性を考慮しない場合の実測危険率

まず，悉皆調査で得られたデータについて測定誤差を考慮しない通常のt検定を行った場合，帰無仮説「母平均の差はゼロ」を誤って棄却してしまう確率（実測危険率）はどのようになるのか，モンテカルロ法を用いたシミュレーションによって考察してみよう。

表１は，シミュレーション用に作成したデータである。ここには，10人の生徒について介入の前（事前）と後（事後）における真値が入力されている。表の最下段に示したように，事後の真値が事前の真値とどの程度強い正の相関

関係を持っているかによって３つの条件（A，B，C）を設定してある。事後の真値はいずれも事前の真値を生徒間で入れ替えて作成したものであるので，平均・分散・標準偏差の値は事前の真値と違いはない。従って，事前と事後の真値の平均値差はいずれの条件においてもゼロであり，「母平均の差はゼロ」という帰無仮説が成り立っている状態を表していることになる。シミュレーションは，事後の真値を事前との相関係数の違いで３水準設定したのに加え，測定の信頼性についてもρ＝.20，.50，.80の３水準を設定した。ただし，事前と事後の測定の信頼性は等しいと仮定した。従って，これら３水準同士の組み合わせで合計９つの条件が設定されたことになる。誤差は平均が０の正規分布に従うと仮定した。測定の信頼性は誤差の分散の値によって操作した。このようにして発生させた誤差の値を用いて，式(1)に従って観測値を得た。得られたデータセットについて，対応のある２群の平均値差に関するt検定を５％水準で実行した。以上の手続きを10,000回繰り返し，「母平均が等しい」という帰無仮説が実際に棄却された割合を示したのが表２である。シミュレーションにはR（R Development Core Team, 2010）を用いた。

表１：事前と事後の真値

生徒	事前	事後		
		A	B	C
1	1	2	2	2
2	1	4	3	2
3	2	1	1	1
4	2	3	2	1
5	3	1	1	3
6	3	5	5	3
7	4	3	4	5
8	4	5	5	5
9	5	2	3	4
10	5	4	4	4
平均	3.00	3.00	3.00	3.00
分散	2.00	2.00	2.00	2.00
SD	1.41	1.41	1.41	1.41
事前との相関		.20	.50	.80

表２：実測危険率（通常のt検定）

事前と事後の相関係数	信頼性		
	.20	.50	.80
.20	.0358	.0076	.0002
.50	.0380	.0195	.0017
.80	.0431	.0339	.0094

ここから，悉皆調査によって得られたデータについて通常のt検定を行うことが問題であるのは明らかである。まず，実測危険率は事前と事後の真値間の相関関係が弱いほど理論値α＝.05よりも小さくなる。しかし，むしろ深刻なのは信頼性が高い場合である。表２からは，測定の信頼性が高くなればなるほど実測危険率が理論値を大きく外れて低くなる様子が見て取れる。しかも，この傾向は極めて顕著であり，信頼性がρ＝.80で真値間の相関係数がρ_{12}＝.20である場合，10,000回の検定のうち棄却されたのはわずか２回にすぎない。検定における第１種の誤りと第２種の誤りが帰無仮説の真偽を軸に表裏の関係にあることを考えると，第２種の誤りを犯す確率は信頼性が高くなるにつれて理論上の値よりも高くなることが予想される。これは，検定力の低下を示しているから，測定の信頼性が高くなるほど実際には偽である帰無仮説を正しく棄却できる確率が低くなることになる。信頼性の高さは測定誤差の少なさを示しているから，信頼性が高いほどデータのばらつきのうち確率的に変動する成分の割合が少なくなり，観測値がすべてサンプリングに伴って確率変動していると仮定している通常の検定を行うと問題が生じることは容易に想像できる。また，データにおいて確率変動する部分が理論上想定されているよりも少ないのであるから，帰無仮説が誤って棄却されなくなる確率が高まるというのも納得のいく結果である。

一般に，測定の信頼性は妥当性が高いための必要条件であり（南風原，2002b）高いに越したことはないとされているから，信頼性が高いという測定上望ましい性質が満たされているほど統計的推測がうまく働かなくなるというのは看過できない問題である。すでに述べたように，悉皆調査であっても観測値を確率的に変動させる要素が存在するならば，統計的推測の対象とすることには意味がある。しかし，それを行うに当たっては観測値のうち確率的に変動する成分の占める割合が測定の信頼性に応じて異なることを考慮する必要がある。少なくとも，上のシミュレーションで扱った対応のある2群の平均値差に関するt検定のような，通常の標本理論にもとづく統計的推測の方法をそのまま用いるべきではない。そうした検定結果に意味があると主張するのは，実態はどうであれその測定には信頼性がほとんどないということを自ら認めているようなものである。信頼性の低い測定によって得られたデータには対象者が本来取るはずである値（真値）の個人差はほとんど含まれていないから，そうしたデータを分析の対象とすること自体が妥当性の観点から問題である。いずれにせよ，悉皆調査で得られたデータに対し通常用意されている統計的推測の手法を適用することが不適切であることに変わりはないと言える。

4　測定の信頼性を考慮した検定

4.1　検定統計量と自由度

以上を踏まえて，ここでは測定の信頼性を考慮した対応のある2群の平均値差に関する検定方法を提案する。まず，以下のようなモデルを考える。

$$x_d = t_d + e_d,\ e_d \sim N(0,\ \sigma_d^2) \tag{4}$$

ただし，x_d，t_dおよびe_dはそれぞれ事前から事後への変化量（差得点）の観測値，真値，および誤差を表す。また，σ_d^2は誤差の分散である。また，観測値間の相関係数をr_{12}とし，2回の測定に共通の信頼性係数ρは既知であるとする。

このとき，差得点の信頼性は

$$\rho_d = \frac{\rho - r_{12}}{1 - r_{12}} \tag{5}$$

となることが知られている（渡部，1993）。

これと信頼性係数の定義を利用して，差得点の観測値の分散V_dから誤差分散σ_d^2を以下のように推定する。

$$\hat{\sigma}_d^2 = V_d \times (1 - \rho_d) \tag{6}$$

これを用いて，検定統計量tを以下のように定義する。

$$t = \frac{\bar{x}_d}{\sqrt{\hat{\sigma}_d^2 / N}} \tag{7}$$

ただし，$\bar{x}_d$は差得点x_dの平均，Nは調査対象者の数である。データから計算された検定統計量の値をt分布における棄却の限界値と照らし合わせて帰無仮説の検定を行うのであるが，ここでは自由度を

$$df = \frac{N-1}{1 - \rho} \tag{8}$$

とする。実は，仮に自由度を通常の対応のある2群のt検定と同様$N-1$とすると，信頼性が高くなるにつれて実測危険率が理論値よりも小さくなるという前節で見られた傾向が，若干残ったままとなってしまう。式(8)は，これを補正するための措置である。式の成り立ちから明らかなように，測定の信頼性が高くなるほど検定の際の自由度が大きくなり，帰無仮説がより棄却されやすくなるように補正が行われている。

4．2　信頼性を考慮した場合の実測危険率

表３は，前節と同様の手続きでシミュレーションを実行した場合の検定における実測危険率を表したものである。ただし，帰無仮説「母平均の差はゼロ」を検定する際には，前項で示した検定統計量（式(7)）および補正後の自由度（式(8)）が用いられている。通常のt検定を実行した場合（表２）と比べて，測定の信頼性を考慮して検定を行った場合は実際の危険率が理論上の値とかなり近くなっていることがわかる。

表３：実測危険率（信頼性を考慮したt検定）

事前と事後の相関係数	信頼性		
	.20	.50	.80
.20	.0491	.0466	.0491
.50	.0533	.0496	.0482
.80	.0503	.0451	.0490

5　まとめ

最後に，本研究で得られた結果について総合的な考察を行っておく。本研究では，サンプリングに伴ってデータが確率的に変動しないと一般的にされている悉皆調査における統計的推測について取り上げた。すでに述べた通り，サンプリングが行われていなくても，測定に伴う誤差が確率的に観測値を左右するのであれば，誤差分散を観測値の分散から切り分けて推定して検定統計量を構成し，補正した自由度を用いることで適切に検定を行うことが可能である。ここでは対応のある２群の平均値差に関する検定しか取り上げなかったが，同様の議論は相関係数の検定やそれ以外の分析における統計的推測においても可能である。ただし，検定の際の自由度（式(8)）からも示唆されるように，悉皆調査における検定は測定の信頼性が完全であれば意味をなさない。これは，$\rho=1$のとき誤差分散σ_d^2が０であることを考えれば明らかである。

本稿では，対応のある２群の平均値の比較という非常にベーシックな手法について，２回の測定に共通の信頼性が既知であるという極めて非現実的なシチュエーションを想定して議論を展開した。こうした想定の下で提案された手法は，明らかに現実のデータ分析に耐えうるものではない。しかし，一方でこれ以上現実的な場面への適用可能性を考慮した手法の提案を行うことの意義はないだろう。というのも，帰無仮説を理論通りの確率で棄却できるような手法をいかに現実的な条件の下で提案したとしても，測定の信頼性が十分である程度に応じて統計的推測が意味をなさなくなるという本質が変わるものではないからである。そもそも測定の信頼性が低ければ妥当性も低いのであるから，そうした質の低いデータをもとに何かを主張しようとすること自体に無理があるといえる。従って，学校評価等で必要とされるデータを得る際には，こうした推測の問題を考えるよりも，いかに妥当性の高い（従って信頼性の高い）測定を行うかということが最も重視されるべきであろう。そのような質の高いデータが得られていれば，統計的推測を行うことの意義自体がほとんど存在しないのである。

ただし，本研究はいわゆる確率化テスト（ランダマイゼーション検定）の意義を否定するものではないことに注意されたい。確率化テストはサンプリングに伴うデータの変動を考慮しているわけではないが，複数の群への割り付けを無作為化する過程が存在することにより母数に関する確率的な議論をすることが可能である（橘，1997）。すなわち，サンプリングに伴う変動がないこと自体が統計的推測の意義をなくすわけではない。このことと本稿で述べた信頼性に関わる議論とを混同すべきではない。

参考文献

Cohen, J.（1988）. *Statistical power analysis for the behavioral sciences*（2nd ed.）. Hillsdale, NJ: Erlbaum.

Gulliksen, H.（1950）. *Theory of mental tests.* NY: Wiley.

南風原朝和（2002a）．モデル適合度の目標適合度－観測変数の数を減らすことの是非を中心に　行動計量学，29，160-166．

南風原朝和（2002b）．心理統計学の基礎－統合的理解のために　有斐閣

R Development Core Team（2010）. *R: A language and environment for statistical computing.* Vienna, Austria.

橘敏明（1997）．確率化テストの方法－誤用しない統計的検定　日本文化科学社

渡部洋（1988）．心理・教育のための多変量解析法入門　福村出版

渡部洋（1993）．心理検査法入門　福村出版

Statistical inference in census

Taichi OKUMURA*

ABSTRACT

Researchers often practice statistical inferences on parameters using census data. In census, data cannot vary stochastically through sampling process. Is it meaningful to practice statistical inference for census data? In this article, the author investigates the meaning of the statistical inference in census from viewpoint of the reliability of measurement. Results of the Monte-Carlo simulation study indicate that some serious problems happen when general paired t-test is applied to the pretest-posttest design census data. The author proposes an alternative t-test statistic and modified degrees of freedom for testing the mean difference considering the reliability of measurement. Nevertheless, such statistical methods get meaningless under high-reliability measurement situation, for randomly varying component decreases. Obtaining high-quality data should be emphasized rather than elaborating statistical inference methods for data analyses of census.

* School Education