情報教育研究における統計的方法の利用

奥村 晴彦*

2012 年 7 月 30 日

概要

情報教育論文での統計的方法の利用実態を調査し，問題点を指摘した。特にアンケートで多用されるいわゆるリッカート尺度のデータを例にとり，統計的方法の望ましい利用法を論じた。

Statistical Methods in IT Education Research

Haruhiko Okumura*

We survey the usage of statistical methods in Japanese IT education research, and discuss how it is improved, taking as an example so-called Likert-type data from questionnaire studies.

1 はじめに

「統計的方法を使わなければ論文でなく随筆だ」
——先輩委員@某ジャーナル編集委員会

米国のコンピュータ関係の学会 ACM（Association for Computing Machinery）の CHI（Human-Computer Interaction）コミュニティで，統計的方法の利用についての議論が行われている。

発端は，ACM の CHI 2010 で Kaptein ほか [1] が前年の CHI の論文を調べ，そのうち 45% がいわゆるリッカート尺度のデータを扱っていながら，8% しか正しい統計的方法（ノンパラメトリック法）を使っていなかったとする報告である。いわゆるリッカート尺度のデータとは，被験者の反応を例えば 5 段階の尺度（例えば「反対」「どちらかといえば反対」「どちらでもない」「どちらかといえば賛成」「賛成」）で尋ねて 1〜5 に数値化することであり，多くのアンケートでこの 1〜5 という数値を間隔尺度（「1 と 2 の隔たりは 2 と 3 の隔たりに等しい」等々）のように扱い，平均を求めたり t 検定したりしている。

この Kaptein ほかの問題提起を ACM の会誌 *Communications of the ACM* の 2012 年 5 月号で Robertson [2] が取り上げ，さらに Kaptein と Robertson [3] は CHI で使われる統計的方法全般に話を広げ，大部分の論文は正しい統計的方法を使っていないと断じた。

彼らのいう「正しい」統計的方法には議論の余地があるが，そのことを含めて，どのような統計的方法を用いるべきかという議論は，過去にいろいろな学術コミュニティで起こっており，そのたびに研究方法が徐々に改善されてきているが，日本の情報教育研究ではまだそのような議論を耳にしない。

本稿では，情報処理学会論文誌情報教育特集号における統計的方法利用の現状を調査し，この分野での統計的方法利用のあるべき形を私見として提示する。異論も多々あることと思うが，シンポジウムでの議論を通じて，方向性を示せれば幸いである。

* 三重大学 (Mie University)

2　日本の情報教育研究における統計的方法の利用の現状

実験や調査を伴う学問分野では，統計的方法（統計的仮説検定，区間推定など）が広く使われている。

例えば CiNii でオープンアクセスできる『教育心理学研究』（日本教育心理学会）の最新巻（Vol. 58, 2010 年）の原著論文 37 編を調べたところ，34 編が何らかの統計的方法[*1] を使っていた。

一方，『情報処理学会論文誌』48 巻 8 号（2007 年），49 巻 10 号（2008 年），50 巻 10 号（2009 年），51 巻 10 号（2010 年），52 巻 12 号（2011 年）の情報教育特集論文をすべて調べたところ，何らかの統計的方法を用いていたのは 32 編のうち 13 編であり，『教育心理学研究』と比べて少ない[*2]。統計的方法ごとの論文数は次の通りである：

- 平均値の差の検定（おそらく等分散の t 検定）5
- 平均値の差の検定（Welch の検定）1
- 分散の検定（F 検定）に続く平均値の差の検定（等分散の t 検定または Welch の検定）1
- Pearson の相関係数の検定（t 検定）4
- 分割表の検定（χ^2 検定）1
- Mann-Whitney の U 検定 1
- 1 元配置分散分析（F 検定）1
- 2 元配置分散分析（F 検定）3
- 共分散構造分析（SEM）のパス係数の検定 1
- 多重比較（Ryan 法）1

統計的方法には疑わしいものもあり，必ずしも論文の価値を増すものばかりではない。しかし，開発した教育法やツールの有効性を示すデータを収集し，正しい統計的方法によってその統計的有意性を示すことは，意味があることであるし，情報教育研究が教育研究として認められるためにも必要なステップである。

以下では個々の問題点について述べる。

[*1] $p < .05$ 等々で p 値を示しているもののほか，MCMC を使って統計量の標準誤差を求めている 1 編も含めた。

[*2] もちろん研究対象が異なるのでこの比較は論文誌の質とは無関係である。

3　統計グラフの問題

ほとんどの論文は表や統計グラフを含んでいる。表はグラフで可視化するほうがわかりやすい。良いグラフはそれだけで統計的仮説検定に勝る可能性がある。良いグラフの要件は他所でも述べたが [4]，特に論文に掲載するグラフとして気づいたことをまとめておく。

■できればベクトル図にする　ビットマップ形式，特に JPEG エンコーディングの図は，高解像度で印刷した際に荒さが目立つことがある。できればベクトル形式にしたい。

■色に頼らない　モノクロ印刷した際に読みづらくなる。

■3 次元グラフは使わない　3 次元グラフは奥行きによって視覚的に誤認を誘発する可能性があるために，特に論文では好ましくないというのが著者の主張である。

■名義尺度の変数の度数分布に折れ線グラフを使わない　これについても異論がありうるが，順序関係がない場合に線で結ぶことは避けたいというのが著者の主張である。

■エラーバーを描く　誤差を含む量を表すグラフには，エラーバー（誤差棒）を描きたい。平均値のグラフでも，平均値の標本誤差に相当する標準誤差（standard error）をエラーバーで表す。データの標準偏差 $s = \sqrt{\sum(x_i - \bar{x})^2/(n-1)}$ をバーで表すこともあるので，標準誤差 $s/\sqrt{n}$ であることを明記すべきである。

4　平均値の差の検定の問題

調査した情報教育関係の論文では，平均値の差の検定（t 検定）が最も広く用いられていたが，対応のある場合の検定か，対応のない場合の検定かが，明確に示されていないものがあった。また，等分散を仮定した検定か，等分散を仮定しない検定（Welch の方法）かが，明確に示されていない論文も多かった。分散が等しい理由がないのであれば，等分散を仮定しない検定をすべきである（R の `t.test()` はこれがデフォルトである）。

さらに問題があるのは，先に等分散かどうかの検定をして，それによって t 検定で等分散を仮定するかどうかを決めるいわゆる 2 段階検定である。古い教科書にそのようなやり方が書かれていたのが原因であろうが，この方法は理論的にも実験的にも問題がある [5][*3]。

5 例：リッカート型データの比較を例として

教育や HCI 分野の論文では，いわゆるリッカート尺度のデータ（例えば「非常に反対」「やや反対」「どちらでもない」「やや賛成」「非常に賛成」をそれぞれ 1〜5 の整数とする）を扱うことが多い。

次の表は，従来型教育と ICT 利用教育に 20 人ずつ振り分けて，対象領域に興味を感じた度合を 5 段階（1〜5）で尋ねた際の度数（人数）を表す架空データである。最下行の midrank については後述する。

段階	1	2	3	4	5	平均
従来型教育	4	5	6	3	2	2.7
ICT 利用教育	1	4	3	6	6	3.6
合計	5	9	9	9	8	
midrank	3	10	19	28	36.5	

5.1 統計ソフトの選択

少なくとも Excel 2007 までの統計関係の機能については，批判が多い [6, 7, 8]。以下では最近注目されているオープンソースの統計・データ解析ツール R を用いる [9, 10]。

5.2 t 検定

上の例にそのまま t 検定（等分散を仮定しない Welch の方法）を適用すれば，次のようになる。`rep(`a`,`n`)` は n 個の a の反復，`c(...)` は結合を意味する。

```
x = c(rep(1,4),rep(2,5),rep(3,6),
      rep(4,3),rep(5,2))
y = c(rep(1,1),rep(2,4),rep(3,3),
      rep(4,6),rep(5,6))
t.test(x,y)
```

結果は $p = 0.03058$ を得る。

ここでこの結果を提示する場合，「$p < .05$ である」と書く習慣が根強い。これは数表で検定を行っていたころの名残りで，現在では「$p = .031$ である」などと報告することが薦められている。例えば米国心理学会のマニュアル [11, p. 114] には次のように記されている：

> When reporting p values, report exact p values (e.g., $p = .031$) to two or three decimal places. However, report p values less than .001 as $p < .001$. The tradition of reporting p values in the form $p < .10$, $p < .05$, $p < .01$, and so forth, was appropriate in a time when only limited tables of critical values were available. ...

ただし，表の中などではアスタリスク (*, **, ***) でそれぞれ $p < .05$, $p < .01$, $p < .001$ を表す習慣は認められている [11, p. 139][12]。

ここで，$p < .05$ を「統計的に有意」とする習慣のある分野が多い（$p < .05$, $p \leq .05$ のどちらが正しいか議論になることもある）。しかし，ボーダーラインの場合に何とか「有意差」を出そうとして，さまざまな統計的方法を試み，$p < .05$ になった方法を採用するという状況は本末転倒である。

検定と p 値については以下の第 6 節でさらに詳しく述べる。

5.3 ノンパラメトリック検定

リッカート型データの t 検定で，「1〜5 の値は間隔尺度の量と見なすことができないので，t 検定は使うべきでない」と言われることがある。このように言われた場合は，いわゆるノンパラメトリック検定を使うのがよいであろう。ノンパラメトリック検定は，順位に基づく検定で，上の例の場合，よく使われるのは Wilcoxon-Mann-Whitney 検定（Wilcoxon の順位和検定または Mann-Whitney の U 検定とも呼ばれる）である。これは，全 40 名を併合したデータで順位を付け，第 1 群の順位の和に基づいて p 値

[*3] ただし，分散が有意に異なるかどうかを調べることが無意味というわけではない。この点をご指摘くださった査読者に感謝する。

を計算する（第 1 群の個数を n として，第 1 群の順位和からその最小値 $n(n+1)/2$ を引いた値 U を使うことが多い)。上のデータ例では，表の最下行の midrank が全 40 名の順位（小さい順）の階級ごとの中央値である。これに第 1 群の人数を掛けて

$$U = 3\times 4 + 10\times 5 + 19\times 6 + 28\times 3 + 36.5\times 2 - 20\times(20+1)/2 = 123$$

となり，それに対応する p 値は

```
wilcox.test(x,y)
```

で $p = 0.03435$ と求められる。ただし，このようにタイ（等しい値）のある場合は近似計算となるので，次のような厳密版を使うほうが正確になる：

```
library(coin)
wilcox_test(c(x,y) ~
  factor(c(rep("x",length(x)),
           rep("y",length(y)))),
  distribution="exact")
```

こちらの結果は $p = 0.03611$ である。

Wilcoxon-Mann-Whitney 検定の統計量 U は，2 群 $\{x_i\}, \{y_j\}$ について $x_i > y_j$ を満たす対の数と $x_i = y_j$ を満たす対の数の半分との和

$$U = n(x_i > y_j) + \tfrac{1}{2}n(x_i = y_j)$$

に等しいので，順位を陽に持ち出さなくても解釈できるが，実質的にはリッカート尺度の得点 1〜5 を順位（midrank）で置き換えた検定にほかならない。リッカート尺度の得点 1〜5 の和に意味がないのであれば，順位の和には意味があると言えるのか。

結局，どのような得点を階級に与えても，その和は正規分布に近づき，検定に使える。その意味で，どちらの方法も可とすべきであると考える。また，順位を使う方法は外れ値に強いという利点があるが，このようなリッカート得点では活かされない。さらに，Wilcoxon-Mann-Whitney 検定は分散の等しい t 検定に相当するもので，分散が異なる場合には正しい結果を与えない。むしろ Wilcoxon-Mann-Whitney 検定よりも，リッカート得点または midrank について，分散の異なる場合の t 検定（Welch の方法）を使う方が望ましい。例えば midrank に t 検定を適用すると次のようになる：

```
x = c(rep(3,4),rep(10,5),rep(19,6),
      rep(28,3),rep(36.5,2))
y = c(rep(3,1),rep(10,4),rep(19,3),
      rep(28,6),rep(36.5,6))
t.test(x,y)
```

結果は $p = 0.03129$ である。

全体の個数を N とすると，midrank は $1 \le r \le N$ の範囲の値をとる。これを $0 < a < 1$ の範囲に規格化するには $a = (r-0.5)/N$ とすればよい。この a を ridit と呼ぶ。これをさらに標準正規分布の累積確率の逆関数で変換したもの $\Phi^{-1}(a)$ を正規得点（normal score）と呼ぶ。これは $\Phi^{-1}(r/(N+1))$ と定義することもある。これらはいずれも元のリッカート得点に置き換えて使えるものである [13]。

なお，Wilcoxon-Mann-Whitney 検定を分散の異なる場合に拡張した方法として，Brunner-Munzel 検定 [14] がある：

```
library(lawstat)
brunner.munzel.test(x,y)
```

この結果は $p = 0.02622$ であり，値が小さいことから，検出力が高い検定であることがわかる。

5.4 相関係数に基づく方法

この例は従来型教育と ICT 利用教育を比較するものであるが，例えば従来型教育，部分的 ICT 利用教育，完全 ICT 利用教育のように，順序関係のある複数の方法を比較することも考えられる。このような順序尺度どうしの関連を見る方法の特殊な場合として，この例を捉えることもできる。

このとき，一方の変数は従来型か ICT 利用かを表す変数である：

```
x = c(rep(1,20),rep(2,20))
```

もう一方は，5 段階のデータである：

```
y = c(rep(1,4),rep(2,5),rep(3,6),
      rep(4,3),rep(5,2),
      rep(1,1),rep(2,4),rep(3,3),
      rep(4,6),rep(5,6))
```

これらの相関係数を

```
cor.test(x, y)
```

で調べれば $p = 0.03128$ となり，t 検定の結果と一致する。一方，ノンパラメトリックな方法，例えば Kendall の τ を使えば，

```
cor.test(x, y, method="kendall")
```

で $p = 0.03319$ となる。

6 検定と多重性

一般的に，統計的仮説検定といえば，データを取得する前にある有意水準 α を定め，$p \leq \alpha$（あるいは $p < \alpha$）であれば帰無仮説を棄却し，そうでなければ態度を保留する（あるいは帰無仮説を「採択」する）という考え方が広く行われている。

これに対して，p 値はいわば確信の度合を表すもので，（例えば $\alpha = 0.05$ として）$p = 0.049$ か $p = 0.051$ かで結論を変える必要はないというのが著者の立場であった。生の p 値を報告すれば，一般的な意味での検定をしたい読者は，自分の α で判断できる[*4]。

しかし，査読者の一人が一般的な立場を擁護されていることもあり，この節の以下では一般的な立場をとり，さらに具体的に $\alpha = 0.05$ として話を進める。

このとき，$p < .05$ であれば統計的に有意とされるが，仮に 20 通りの比較をして p 値を 20 個求めれば，全くの偶然でもそのうち（期待値の意味で）一つは $p < .05$ となる。このような多重性に対処するために，n 通りの比較を行うならば，基準を $p < .05$ でなく $p < .05/n$ とする（Bonferroni の方法）といったことが行われている。例えば 10 個の変数があれば $n = {}_{10}C_2 = 45$ 通りの比較が可能であり，Bonferroni の方法を採用すれば $p < .05/n \approx 0.001$ でないと有意にならない。さらに仮定を置いて，より有意になりやすくした方法が，いくつも考えられている。

[*4] 多重比較についても生の p 値を報告すれば読者は自分の好きな方法で多重性の考慮をすることができる。

7 おわりに

情報教育論文での統計的方法の利用実態を調査し，その陥りやすい問題点を論じた。特にアンケートで多用されるいわゆるリッカート尺度のデータを例にとり，どのような統計的方法を使えばよいかを論じた。本稿をきっかけに情報教育研究コミュニティで統計的方法の使い方について議論が行われることを期待する。

最後に，たいへん詳しく見てくださった匿名の査読者に感謝する。

参考文献

[1] Maurits Kaptein, Clifford Nass, and Panos Markopoulos, “Powerful and Consistent Analysis of Likert-Type Rating Scales,” *Proceedings of the 28th International Conference on Human Factors in Computing Systems: CHI 2010*, 2391–2394 (2010).

[2] Judy Robertson, “Likert-type Scales, Statistical Methods, and Effect Sizes,” *Communications of the ACM*, **55**, No. 5, 6–7 (2012).

[3] Maurits Kaptein and Judy Robertson, “Rethinking Statistical Analysis Methods for CHI,” *Proceedings of the 30th International Conference on Human Factors in Computing Systems: CHI 2012*, 1105–1113 (2012).

[4] 奥村晴彦「情報教育と統計」，情報処理学会研究報告「コンピュータと教育」2008-CE-97（情処技報 Vol.2008, No.128），81–88（2008）.

[5] 奥村晴彦「2 段階 t 検定の是非」http://oku.edu.mie-u.ac.jp/~okumura/blog/node/2262/

[6] B. D. McCullough and David A. Heiser, “On the accuracy of statistical procedures in Microsoft Excel 2007,” *Computational Statistics and Data Analysis* **52**, 4570–4578 (2008).

[7] A. Talha Yalta, “The accuracy of statistical distributions in Microsoft® Excel 2007,” *Computational Statistics and Data Analysis* **52**, 4579–4586 (2008).

[8] B. D. McCullough, “Microsoft Excel’s ‘Not The Wichmann-Hill’ random number generators,” *Computational Statistics and Data Analysis* **52**, 4587–4593 (2008).

[9] R Development Core Team, “R: A Language and Environment for Statistical Computing”, http://www.R-project.org/

[10] 奥村晴彦「R を使った情報教育」情報処理学会情報教育シンポジウム SSS2010 論文集 (2010).

[11] *Publication Mannual of the American Psychological Association, 6th ed.* (American Psychological Association, 2010).

[12] Leland Wilkinson and the Task Force on Statistical Inference, APA Board of Scientific Affairs, “Statistical Methods in Psychology Journals: Guidelines and Explanations”, *American Psychologist* **54**, 594–604 (1999).

[13] Alan Agresti, *Analysis of Ordinal Categorical Data, 2nd ed.* (Wiley, 2010).

[14] Edgar Brunner and Ullrich Munzel, “The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation”, *Biometrical Journal*, Vol. 42, pp. 17–25 (2000)