

統計と情報教育研究

奥村晴彦

三重大学教育学部

情報教育と統計教育を扱うシリーズ最終回として、(1) データで読む日本の情報教育環境, (2) 情報教育研究と統計的方法, (3) オープンデータ, オープンリサーチ, リプロデュースブルリサーチの話題を扱い、最後に全体をまとめる。

データで読む日本の情報教育環境

2003年・2006年のPISA (Programme for International Student Assessment, 生徒の学習到達度調査)^{☆1} (OECD (Organisation for Economic Co-operation and Development, 経済協力開発機構)^{☆2} 生徒の学習到達度調査) で日本の成績が悪かったという「PISAショック」は、ゆとり教育転換の一因となった。また、知識型の学力と異なる「PISA型学力」、たとえば統計グラフから意味を読みとる問題などが注目され、これが中学数学の「資料の活用」の全学年での復活や、必修の高校「数学I」の「データの分析」につながったのかもしれない。2015年の大学入試センター試験の数学では「データの分析」の内容がしっかり出題されていた。一方で、統計分野をあまり勉強する機会がなかった先生方も多く、教育現場での戸惑いも見られる。

PISAにはアンケート調査もある。生徒のICT (情報通信技術) 利用に関する項目も多い。PISAのサイト^{☆1}には、回答の個票データと、それを読むためのSAS・SPSSプログラムが公開されている。データは

固定フォーマットのテキストであるが、簡単にCSV化できる。筆者のWebページ^{☆3}にCSV化のRubyスクリプトと、例としてPISA2012データをCSV化したものがある。これは、高校1年相当485,490人(うち日本は6,351人) × 634項目の巨大な個票データで、R標準のread.csv()では時間がかかる。readrパッケージのread_csv()なら数10秒で読めた。

いったんオンメモリに読み込めば、集計は瞬時にできる。たとえば図-1は、学校外(自宅など)でコンピュータを利用して宿題を行うことが「まったくか、ほとんどない」と答えた生徒の割合である。ただし、データなし(N/A)・無効(Invalid)・無答(Missing)は分母から除いてある。

この図から分かるように、日本では宿題にコンピュータを使わない生徒(コンピュータを使う宿題を出さない先生)が圧倒的に多い。

学校外でのコンピュータを使った活動についての問いは、これを含めて7問あるが、「勉強についてほかの生徒とメールする」以外はすべて、「まったくか、ほとんどない」のトップは日本である。

また、図-2から分かるように、勉強に限らず自宅でコンピュータ(デスクトップまたはノート)を使う割合では、日本は下位であるが、ネット接続できる携帯電話(スマホなど)の利用者は多い。

つまり、諸外国と比べて日本では、生徒は学校外で(携帯やスマホ以外の)コンピュータを使う機会が非常に少ない。学校で教材として使うことはあっても、

☆1 <http://www.oecd.org/pisa/>

☆2 <http://www.oecd.org/>

☆3 <http://oku.edu.mie-u.ac.jp/~okumura/stat/pisa.html>

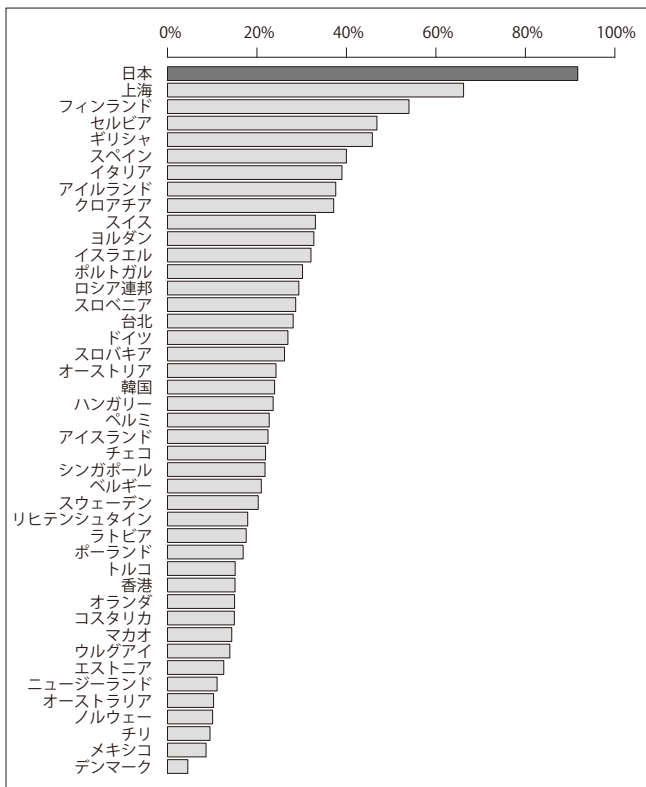


図-1 PISA2012 生徒アンケート “How often do you use a computer for the following activities outside of school? – Doing homework on the computer.” (IC09Q06) に “Never or hardly ever” と答えた生徒の割合

学校外で知的生産ツールとして使うことは少ない。このことは、日本の情報教育を考える上で重要なポイントになる。

ほかにも PISA のデータからはいろいろなことが読みとれる。また、PISA 以外にも、PIAAC (Programme for the International Assessment of Adult Competencies, 国際成人力調査) ^{☆4}, TIMSS (Trends in International Mathematics and Science Study, 国際数学・理科教育調査) ^{☆5}, PIRLS (Progress in International Reading Literacy Study, 国際読書力調査) ^{☆5}, TALIS (Teaching and Learning International Survey, 国際教員指導環境調査) ^{☆6} など、いろいろなデータが公開されている。これらを集計し、そこから意味を読みとることは、学生の課題としても有意義であろう。

☆4 <http://www.oecd.org/site/piaac/>

☆5 <http://timssandpirls.bc.edu/>

☆6 <http://www.oecd.org/edu/school/talis.htm>

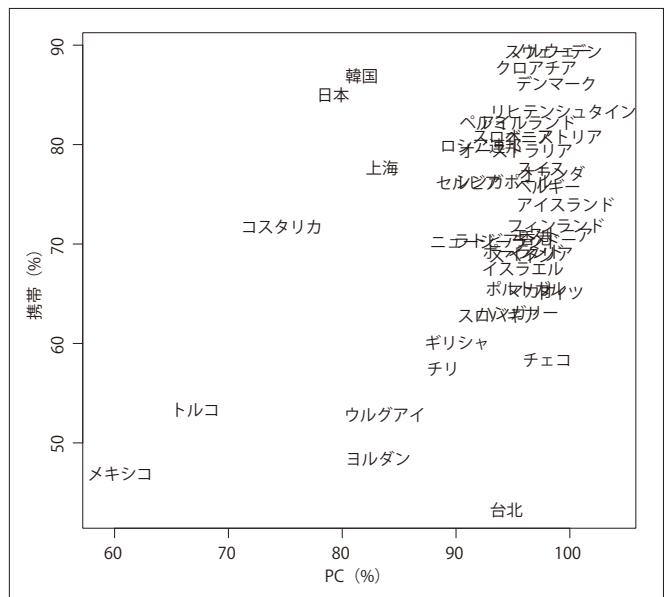


図-2 PISA2012 生徒アンケート “Are any of these devices available for you to use at home?” (IC01) で (a) Desktop computer または (b) Portable laptop, or notebook のどちらかまたは両方に “Yes, and I use it” と答えた割合を横軸、(g) <Cell phone> (with Internet access) に “Yes, and I use it” と答えた割合を縦軸にとった。自宅でのコンピュータについての問いは ST26Q04 にもある

ちなみに、PIAAC の結果では、16 ~ 29 歳の「コンピュータ経験なし」「ICT コア不合格」「コンピュータ調査拒否」を合わせた割合は OECD 国中で日本がトップ(つまり最悪)である。

なお、図-2 のように国ごとの平均値の散布図を描くと、いろいろな相関関係が見えるが、交絡 (confounding) によることが多く、安易に因果関係を推測することは禁物である。

情報教育研究と統計的方法

PISA などの大規模調査のデータであればほとんど問題にならないが、我々が研究で扱う大部分のデータは少数であり、標本誤差の問題を避けて通れない。そのため、統計的検定や区間推定といった統計的方法が用いられる。

情報教育論文でどのように統計的方法が用いられているかを調査したことがある^{1), 2)}。その後、本会論文誌「教育とコンピュータ」(TCE) も発刊され、統計的方法の使われる機会も増えた。ここで改めて情報教育論文のための統計について問題提起しておく。あく



までも議論のためであり、異論は多々あることと思う。

本会のジャーナル類が完全電子化され、色が使えるようになったが、色だけに頼るグラフはなるべく避けるべきである。特に赤と緑は区別しにくい人が多い。そのほか、従前から言われていたこととして、3次元グラフは避ける。棒グラフは0から始める。順序のない項目(名義尺度)を横軸とする折れ線グラフは避ける。標本誤差が無視できない場合は、グラフにエラーバー(通常は95%信頼区間)を付ける。さらには、平均値や中央値だけでなく、個々の値をそのままプロットするほうがよいことが多い³⁾。また、論文に載せる図はビットマップではなくベクトル形式が望ましいことが多い。

以下では、架空の例を使って、統計的方法の適用の問題例を示したい。

例 ある教育システムを開発した。その効果を客観的に示すため、当該システムを使った一連の授業の前後に同じテストを行い、対応のある場合の t 検定により、 $p < 0.05$ で有意であることを示した。

何が「有意」なのか。「一連の授業の前後で平均点がまったく同じ」という(明らかに正しくない)帰無仮説が5%水準で棄却されたと言いたいのだろうが、ほとんど情報量がない。より情報量のあるのは、たとえば100点満点(標準偏差15点)のテストの得点が平均10点上昇した(95%信頼区間[8,12])といった書き方であろう。つまり、「有意」かどうかより、具体的な値とその信頼区間を書くほうがよい。もし仮に信頼区間が[-1,21]のように0をはさむなら、従来の言い方をすれば「有意な変化はなかった」。しかし、有意でなくても、信頼区間[-21,1]と[-1,21]とでは印象がずいぶん違う。さらに、この効果がほかのシステムと比べてどれくらい大きいかを示す必要がある。被験者を対照群と実験群に無作為に分けて比較するのが一般的であるが、システムがらみの教育実験なら、A/Bテストのようにランダムに対照教材・実験教材のどちらかを提示して比較する方法もある。

例 システムAの効果とシステムBの効果比べたい。

まず分散が等しいか検定し、等しければ等分散を仮定した t 検定、等しければ等分散を仮定しない t 検定(Welchの検定)をする。

分散が等しいという帰無仮説が棄却できなくても、分散が等しいことが示されたわけではない。分散が等しいかどうか分からなければ、等分散を仮定しない t 検定(Welchの検定)を最初から使うべきである。シミュレーションによる結果も後者だけのほうがよい。

例 「とても」「少し」「あまり」「全然」のような段階で答える回答を4:3:2:1のように点数化して平均を求めたところ、査読者に「等間隔ではない量の平均に意味はない」と言われた。

教育研究で頻出するいわゆるリッカート(Likert)尺度のデータであるが、査読者からノンパラメトリック検定をせよと言われた場合は、 t 検定の代わりにWilcoxon-Mann-Whitney検定(両群の分布が同じと仮定できる場合)またはBrunner-Munzel検定(より一般の場合)を使えばよい。しかし、中心極限定理のおかげで、基の分布や等間隔かどうかにかかわらず、平均値の分布は正規分布で近似できるので、 t 検定(や t 分布に基づく信頼区間)を使っても、大した違いはない。それに、この場合も、有意かどうかよりも具体的な平均値(と信頼区間)のほうが情報量があり、分かりやすい。

例 複数の検定をしたところ、多重比較の補正をせよと査読者に言われた。

20個の検定をしたら、まったく効果がなくても、 $p < 0.05$ となるものが期待値として1個ある。これではまずいので補正せよという話になるのだが、補正の流儀はたくさんあり、たとえばBonferroni法とBenjamini-Hochberg法とではまったく意味が違う。むしろ単純に、いくつの変数を調べたか分かる書き方をした上で、無補正の p 値(あるいは信頼区間)を示し、読者の判断にゆだねるといっているのでどうだろう。

例 システムの効果が上位群と下位群とで違うように感じたが、有意にならないので、上位・中位・下位に分け、それぞれの人数を加減しているうちに、有意になった。

恣意的に群に分けるのではなく、たとえば成績 x とシステムの効果 y の間に線形の関係があることを示したいのであれば、 $y = bx + \varepsilon$ という回帰モデルをあてはめて、 b が 0 でないかどうか検定する（さらには b の信頼区間を求める）。

オープンデータ，オープンリサーチ，リプロデューシブルリサーチ

ソフトウェアで成功を収めたオープンソースの考え方を一般の著作物に拡張した Creative Commons は、MIT (Massachusetts Institute of Technology, マサチューセッツ工科大学)^{☆7} などの Open CourseWare (OCW) や Wikipedia など多くの著作物で採用されている。オープン化の流れはデータの世界にも広がり、データを機械可読な形式・自由なライセンス (CC0, CC BY など) で公開する「オープンデータ」の考え方が定着してきた。上述の PISA のデータもオープンデータの一例である。

オープンデータは、オープンだけでなく、データとして再利用が容易でなければならない。PDF や「ネ申 Excel」では困る(前回参照)。

研究の根拠となるデータを論文とともに公開する動きも出てきた。一部の論文誌ではデータの開示が義務付けられるようになった。開示義務がなくても、figshareなどのサイトを使って自発的にデータを公開する研究者が増えた。figshare は無償で容量制限もなく、論文と同様な識別子 DOI (Digital Object Identifier) も付けられる。

さらに、データだけでなく、その解析コードも公開し、研究の流れ全体を再現可能(reproducible) にしようとする動きも出てきた。そのためのツール類も作られているが、Unix の「make」でもかまわない。データと

解析コードの入ったディレクトリで make と打ち込めば、Makefile に書かれた依存関係に従って(必要なら解析コードをビルドして) データを処理し、R や gnuplot でグラフを作成し、さらには LaTeX で論文をタイプセットする。データや解析コードに間違いが見つければ、修正してもう一度 make すると、影響を受ける部分だけ処理を自動実行できる。

これらの解析コード類や Makefile をまとめて公開すれば、読者はデータから論文までのすべてのステップを検証できる。データに疑問を持った読者は、データだけ自分のものに置き換えて make し直すことができる。

あるいは、専用のツール、たとえば Sweave や knitr を使えば、ちょうど Knuth の「文芸的プログラミング」のように、解析コードとドキュメントを1つのファイルにまとめたものから、解析結果(図表あるいはインラインの数値など)を埋め込んだ論文が生成できる。

3 回の連載のしめくくり

基本データ型がベクトルな言語 R、それを用いたプログラミングとデータ解析、統計的方法、それらの情報教育および情報教育研究への応用、データ再利用の妨げとなるネ申 Excel 退治、オープンデータ、オープン&リプロデューシブルリサーチ——こういったものをまとめれば、教育のネタ、研究のネタになりそうだということで、ずっと考えてきたことをざっくばらんに3回に分けて披露した。ご意見をいただければ幸いです。

参考文献

- 1) 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 (2008).
- 2) 奥村晴彦：情報教育研究における統計的方法の利用，情報処理学会情報教育シンポジウム SSS2011 論文集，pp.91-96 (2012).
- 3) Tracey, L. W., Natasa, M. M., Stacey, J. W. and Vesna, D. G. : Beyond Bar and Line Graphs : Time for a New Data Presentation Paradigm, PLOS Biology, doi:10.1371/journal.pbio.1002128 (2015).

(2015年6月1日受付)

奥村晴彦 (正会員) okumura@okumuralab.org

三重大学教育学部教授(情報教育)。LHA データ圧縮アルゴリズムの開発者、『LaTeX2e 美文書作成入門』(現在第6版)の著者。

☆7 <http://web.mit.edu/>

