

情報教育と統計

奥村晴彦*

三重大学教育学部

情報の科学的な捉え方の教育としての情報教育のあり方が模索されている。特に数的情報はコンピュータ処理に馴染みやすく、表計算ソフトを使った統計計算やグラフ作成が情報教育の一つとして広く行われている。一方で、メディアや論文は正しくない統計データの扱いで溢れている。本稿では、情報教育に関連した統計データの扱い方の問題を論じる。

Informatics Education and Statistics

Haruhiko Okumura*

Faculty of Education, Mie University

In designing informatics education that aims at teaching scientific understanding of information, we find numerical information particularly suitable for computer processing, e.g., statistical computation and visualization by spreadsheet software. Seeing that our society is flawed by unscientific use of statistical data, we set out to discuss the possibility of teaching statistical methods in the context of informatics education.

1 はじめに

情報教育で表計算ソフトによる簡単なデータ解析・グラフ作成の実習をすることは多い。しかし、この背景となるグラフ作成のルールや統計データの解釈のしかたについて、中学・高校の現行学習指導要領の数学ではほとんど教えられていない。

数値データを視覚化し、意味を読み取ることは、情報社会で広く求められる能力であるが、この能力を十分身につけていない社会人が増えるならば、大きな問題である。

実際、われわれの目にする統計グラフの描き

方にルール違反がよくあることが指摘されている [1, 2]。そのようなグラフは、高校の「情報」教科書の類にも見られる。グラフにとどまらず、マスコミなどで目にする統計データの扱い方には、問題のあることがよくある。

この現状を考えると、情報教育において、単に表計算ソフトの操作を教えるだけでなく、そのグラフが適切であるか、さらにはそこからいかに正しい結論を導くかを学ぶことは意義のあることである。単に数学で学んでいないことの補完ではなく、情報デザインの観点からグラフを扱い、メディアリテラシーまたはモデル化とシミュレーションの観点から統計的方法を扱うことができるのは、情報教育ならではのことである。この主張が本稿の一つの柱である。

* okumura@edu.mie-u.ac.jp

本稿のもう一つの柱は、大学の情報教育課程における統計教育、およびわれわれ情報教育研究者の道具としての統計的方法の利用スキルについてである。特に、情報教育を志す学生は、単に情報技術を学ぶだけでなく、教育学の研究法として重要な位置を占める統計的方法を学ぶことにより、情報教育研究において統計的手法を正しく適用する能力を身につけることができる。

情報教育の枠内で統計的な考え方を教えることは、すでに何度か提案されている（例えば大西 [3], 大貫 [4, 5, 6], 垣花 [7], 成田 [8]）。例えば大貫は高校 1 年の授業で日本文教出版『新・情報 A』（2007 年度版）の「クラスの実態を調査し、分析しよう」という単元（8～9 時間）に基づき、統計的検定（ 2×2 分割表の Fisher の正確検定）まで扱っている。日本における統計教育の現状については木村ら [9] が詳しく論じている。

2 学習指導要領における統計

現行学習指導要領（小・中学校は 1998 年、高校は 1999 年告示）では、小学校算数の 2 年で表・グラフ、3 年で棒グラフ、4 年で折れ線グラフ、5 年で円グラフ・帯グラフ、6 年で平均が扱われるが、中学校数学では統計はまったく扱われなくなっている（後述のように 2008 年告示の次期学習指導要領で復活することになった）。

高等学校では、必履修でない数学 B に「統計とコンピュータ」、数学 C に「確率分布」「統計処理」が含まれており、これらを合わせた内容には

- 度数分布表・相関図
- 代表値・分散・標準偏差・相関係数
- 二項分布・正規分布
- 母集団・標本・統計的な推測・推定

が含まれる。しかし、いずれも選択内容であり、実際にはほとんど教えられていない。数学基礎には「身近な統計」という内容があるが、これも履修率はきわめて低い。これ以外に統計が学習指導要領に現れるのは公民と地理歴史だけである（公民では「統計などの資料の見方やその意味、情報の検索や処理の仕方、簡単な社会調査の方法などについて指導するよう留意すること」と記されている）。

新学習指導要領（2008 年告示）では、小学校 6 年で扱っていた「平均」が 5 年に移動し、6 年に「資料の平均や散らばりを調べ、統計的に考察したり表現したりすることができるようにする」という内容が加えられる。中学校数学でも統計的内容が復活し、1 年で「目的に応じて資料を収集し、コンピュータを用いたりするなどして表やグラフに整理し、代表値や資料の散らばりに着目してその資料の傾向を読み取ることができるようにする」、2 年で「不確定な事象についての観察や実験などの活動を通して、確率について理解し、それを用いて考察し表現することができるようにする」、3 年で「コンピュータを用いたりするなどして、母集団から標本を取り出し、標本の傾向を調べることで、母集団の傾向が読み取れることを理解できるようにする」とされている。これらの内容が現行指導要領では抜け落ちていることから、当分は次の事項について学習していない生徒・学生が多い状況が続く。

- 度数分布・ヒストグラム・相対度数・階級
- 代表値（平均値・中央値・最頻値）
- 散らばり・範囲
- 母集団・標本

新指導要領でこれらが復活しても、中学校での扱いには限界がある。例えば「散らばり」の度合としては「範囲」（最大値と最小値の差）が

あるだけで、標準偏差は現れない（ほとんど履修されていない高校の現行数学 B の選択内容に含まれる）。

標準偏差は「偏差値」の意味を理解するためにも必要なはずである。また、「相関がある」という言葉はマスコミでもよく使われるものであり、その正しい意味（相関関係は因果関係でないことなど）を理解することはメディアリテラシーの観点からも大切である。

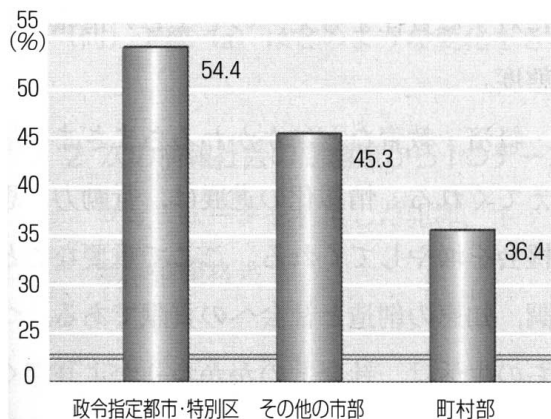
これに対して、高校「情報」では、豊富な実習時間と「表計算ソフト」という武器を活用することにより、数学とは違う切り口でグラフ作成や統計的な考え方を学ぶことができる。表計算については情報 A・C で言及されており、情報 C には「表計算ソフトウェアなどの簡単な統計分析機能やグラフ作成機能などを扱う」とある。教科書によって扱い方の軽重はさまざまであるが、例えば一橋出版『情報 A』は表計算ソフトに 35 ページも割いている。一方、情報 B には「モデル化とシミュレーション」があり、ここで実際に乱数を使って数値実験をすることができる。教科書によっては散布図や相関係数もここで扱っている。

3 統計グラフの描き方の教育

前節でも触れた問題のある統計グラフで、特に多いのは図 1 のように高さが数値と比例しない棒グラフ、図 2 のように必要もないのに 3 次元化して目盛りを読みにくくしたグラフ、3 次元化して特定の部分を強調した円グラフ、凡例との対応を色だけに頼った色覚異常者泣かせのグラフである。

グラフを描く際にはデータの水準（名義尺度 < 順序尺度 < 間隔尺度 < 比例尺度）に注意しなければならない。

- 棒グラフは比例尺度の量を表すのに用い、



総務省「情報通信白書」(平成14年版)より作成
 図10 インターネット利用率 地域別 (平成13年)

図 1 実教新版『情報 A』p. 135 (2004)

情報機器の操作がわからないときにどうしますか

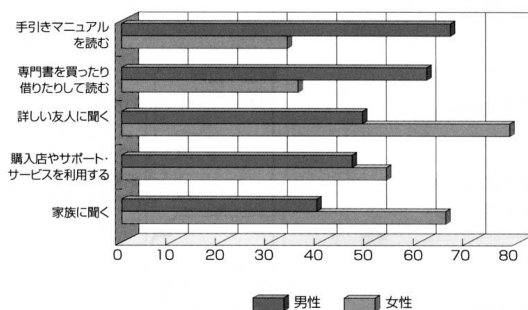


図 2 日本文教出版『情報 A』p. 97 (2005)。斜めから見ているので目盛りが正確に読めない。横軸の単位がわからない。凡例との対応を色だけに頼っている。

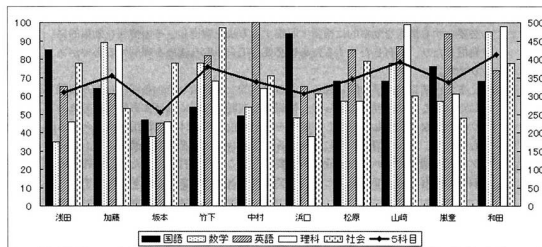


図 3 第一学習社『情報 A 整理と実習』p. 108, 『情報 C 整理と実習』p. 107 (2008)。名義尺度に対して折れ線グラフを使っている。背景が灰色でグラフが見にくい。

棒の長さ（面積）が量に比例するように描くのが基本である。一つだけずば抜けて大きい値があるときに棒の上のほうで省略線を用いることはあるが、図1のように下を省略して全体の長さを切り詰めることは好ましくない。なお、例えば気象観測データについては、雨量は比例尺度であるので棒グラフで表してもよいが、気温は（絶対温度でない限り）間隔尺度であり、棒グラフは使えない。

- 折れ線グラフは、両軸とも間隔尺度以上であるのが基本である（パレート図や Zipf の法則の図示などで順位を間隔尺度のように扱うことはある）。0 点から始める必要はない。特に時系列データについてよく使われる。図3のように名義尺度に対する量を折れ線グラフにするのは異様である。
- 円グラフは全体における割合を表すにはよいが、全体との比較ではなく個々の値どうしの比較には棒グラフがよい。いわゆる3次元(3D)円グラフは、錯覚を利用して特定の部分を大きく見せるためのもので、用いてはならない。
- 色分けして凡例を付けるのは色覚異常者に優しくないのを避けた。凡例と見比べる必要のないグラフが好ましい。
- いわゆるチャートジャンク（情報量のない視覚的な飾り）は情報デザインの観点からも避けるべきである [10]。
- いくつかの度数分布を比較するのに便利な Tukey のボックスプロット（箱ひげ図） [11] や、名義尺度に対する間隔尺度を表す Cleveland のドットプロット [12, 13] も、よく使われるようになった。これらも含めて、どのようなグラフが適当かを選ぶことができるようにしたい。

4 統計のウソとメディアリテラシー

「街で○人に聞きました」という類の統計は、統計誤差だけでなく系統誤差（バイアス）が無視できない。アンケートそのものも、誘導尋問的なものが多い。世論調査の結果も、新聞社によってかなり異なることが多い。

調査としては正しくても、メディアの扱いが間違っていることはよくある。この類の話は、古くはロングセラー『統計でウソをつく法』 [14]、最近では谷岡 [15] など、多くの本で扱われている。特徴的な例として、例えば日々の運動と健康に相関関係が見いだされたときに、「運動すれば健康になる」という因果関係を安易に結論しがちである。しかし、実際は因果関係が逆（健康だから運動する）かもしれないし、あるいは別の要因（例えば経済的余裕、あるいは何らかの遺伝子）が運動・健康の両方の原因かもしれない。

学術論文も同様である。最近では、共分散構造分析のソフトを使い、因果関係を示す矢印の図を描いて計算し、係数が有意に0と異なれば、あたかも因果関係が証明されたかのような錯覚に陥りがちであるが、証明されるのは相関関係だけであることに注意しなければならない。

ただ、メディアリテラシーだけでは統計誤差は評価できない。そこで一般に使われるのは正規分布に基づく統計理論であるが、これは情報教育に馴染まない上に、正規分布が仮定できないことも多い。そこで、情報Bの「モデル化とシミュレーション」を援用する。

5 「モデル化とシミュレーション」による検定と推定

乱数を用いたシミュレーションとしては、円周率をモンテカルロ法で求めることや、待ち行

列のシミュレーション、ランダムウォークなどが従来からよく使われてきた。ただ、いずれも数学寄りの話題であり、必ずしも生徒の興味を惹かない。

そこで、より現実的な問題のシミュレーションを通じて、統計的仮説検定の原理を学ぶことができることを示す。

例えば硬貨を8枚投げて、表・裏の枚数の差が4以上になる確率 $P(6,2) + P(7,1) + P(8,0) + P(2,6) + P(1,7) + P(0,8)$ を当て推量で答えさせてから、実際に表計算ソフトでシミュレーションし、結果を比べる。これと同値の問題として、8人に質問して6人が賛成、2人が反対と答えたとき、本当は賛否の確率が等しいという帰無仮説の下に、賛否の差が観測された差以上(4以上)になる確率(有意確率、 p 値)を求める問題がある。もちろん数学で2項分布を使えば正確な確率が求められるが、シミュレーションでおおよその答えを出すことができる。具体的には、`=RANDBETWEEN(0,1)` と入力し、右にフィルして8個にし、その右に合計を求める。この9個のセルを下にフィルし、最右列の度数分布を求める。それが6, 7, 8, 2, 1, 0のどれかである割合を求めれば3割ほどになる。有意確率 $p \doteq 0.3$ は小さくないので、観測された8:2は珍しい事象ではないことが示された。統計的に有意であるためには概ね $p \leq 0.05$ が条件とされることが多いが、0.05という値に根拠はない。

より高度な、統計学で並べ替え検定(permutation test)と呼ばれるノンパラメトリックな(正規分布を仮定しない)検定も、Excelを使えば簡単にできる(図4)。ここでは、男子5人、女子6人のアンケートの回答(1~5で数値化してある)の平均の差が有意かどうか検定している。少人数であり、正規分布でもないので、通常の t 検定は使えない。そこで、男女に差が

ないというモデル(帰無仮説)を仮定し、男女合わせた11人を乱数で並べ替え、上5人と下6人の平均の差を多数(ここでは1000個)自動生成して、それが実際に起きた差(1.8667)以上になる割合を求める。この例は実際には ${}_{11}C_5 = 462$ 通りの組合せしかないので、すべての場合を調べれば正確な値 $p = 0.02814$ が求められるが、一般には組合せの数は非常に大きいので、乱数によるシミュレーションが適当である。

上の並べ替え検定は非復元抽出であるが、2標本の平均の差の信頼区間を求めるシミュレーション(ブートストラップ)では、各標本から元と同数だけ復元抽出する。平均の差(今度は絶対値をとらない)の分布を上と同様に求め、PERCENTILE()関数で2.5%点と97.5%点を求めれば、95%信頼区間が得られる。Christie [16]は同様な例をいくつか挙げている(ただし並べ替え検定の p 値の計算にはRANK()でなく図4のようにCOUNTIF()を使うのが正しい)。

別の例として、ヒストグラムを描く練習を兼ねた中心極限定理の実験が考えられる。一樣乱数RAND()を3個加えた時点で釣り鐘型になり始め、12個加えて6を引けば標準正規分布 $N(0,1)$ とほぼ寸分違わぬ分布になる。

残念なことに、Excelの乱数は専門家からは信頼されていない。Excel 2003未満の乱数は十分ランダムではなく、Excel 2003初期版はRAND()にときどき負の値が混じるバグがあり、パッチを適用したExcel 2003以降も、正しく実装されていない証拠がある[17]。実務としての統計計算やグラフ作成のためには、オープンソースのR [18]を初め多数の良質なソフトが存在するので、それらを選ぶべきである。

	A	B	C	D	E	F	G	H
1	番号	元データ		乱数	番号	並べ替え		1.867
2	1	5		0.484	5	4		0.767
3	2	3	男	0.399	6	2		2.167
4	3	4		0.113	10	3		1.133
5	4	5		0.603	3	4		1.5
6	5	4		0.865	1	5		1.133
7	6	2		0.294	8	3		1.5
8	7	1	女	0.545	4	5		0.7
9	8	3		0.218	9	1		1.067
10	9	1		0.395	7	1		0.4
11	10	3		0.038	11	4		0.767
12	11	4		0.768	2	3		1.433
13								1.433
14		4.2	男平均		3.6			0.7
15		2.3333	女平均		2.83333			0.333
16		1.8667	差の絶対値		0.76667			1.067
17								0.4
18		0.03	p値					0.333
19								0.4
20								0.767

図4 Excelによる並べ替え検定。列Bに元データを入れ、列Dの乱数の順位を列Eに求め、VLOOKUPで並べ替えたデータを列Fに求める。列B、Fの男女別平均の差の絶対値をそれぞれB16、F16に求める。セルH1に=B16、セルH2に=F16を入れて、G2:H1000を選択し、データテーブルを設定する（「行の代入セル」は空欄、「列の代入セル」は適当な不使用セルにする）。再計算法は「データテーブル以外自動」または「手動」にしておくと重くならない。p値は=COUNTIF(H:H,">="&H1)/COUNT(H:H)で求められる。

6 情報教育研究と統計

実験や調査を伴う学問分野では、統計的方法（特に仮説検定）が広く利用されている。しかし、情報教育研究においては、統計的方法の利用は比較的少ない。

例えば『教育心理学研究』（日本教育心理学会）掲載論文では、56巻1号（2008年）11編（「展望」を除く）のうち9編、56巻2号（2008年）12編のうち10編が、何らかの統計的検定を行っていた。

一方、『情報処理学会論文誌』48巻8号（2007年）の情報教育特集では巻頭言を除く8編のう

ち5編、49巻10号（2008年）の情報教育特集では巻頭言を除く9編のうち4編が何らかの統計的検定を用いている。SSS2008論文33編では6編に過ぎない。

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。

以下は情報教育研究での統計的方法の利用についていくつか気付いたことを列挙しておく。

- 統計ソフトの選択は重要である。定評のあるものを使うべきである。
- 「授業前にプリテスト、授業後にポストテストを行い、対応のある場合のt検定をしたところ、 $p < 0.05$ で有意差が見られた」といった検定をしばしば見るが、テストの内容に直接かかわる授業をした場合、ポストテストの点数のほうが高いのは当然である。点数の差の変動の要因を調べたり、他の学習方法と比較したりするといった実質的に意味のある部分に統計的方法を使うべきである。
- $p > 0.05$ か $p \leq 0.05$ かどうかだけに意味があるのではない。 $p \leq 0.05$ と書くスペースがあるなら $p = 0.012$ のように具体的な値を書く ($0.01 < p \leq 0.05$ を星印*で表すといった略記法はかまわない)。
- 両側検定か片側検定か明確にする（一般には両側検定を使う)。
- 数値を上位群と下位群に分けるためだけに使うのは、数値の情報の大部分を捨てることになり、もったいない。

- 2 標本を比較する t 検定では、等分散を仮定する方法以外に、等分散を仮定しない Welch の方法がよく使われる。等分散である必然性がない場合は後者を使うべきである (R の `t.test()` は後者がデフォルトである)。なお、実験後に等分散の検定をして、その結果によってどちらの t 検定を使うかを定める方法は、多くの統計書で推奨されているが、正しい方法ではない。
- ノンパラメトリック検定、並べ替え検定、ブートストラップ、種々の正確な方法を活用すべきである。
- ノンパラメトリック検定は必ずしも分布を仮定しない検定ではない。例えば中央値が 0 であることを調べる Wilcoxon の符号つき順位検定は分布が 0 を中心に対称であることを仮定しているし、Wilcoxon-Mann-Whitney 検定は 2 標本の分布が等しいことを仮定している。分布が異なる場合 (等分散でない等の場合) は Brunner-Munzel 検定 [19] などを用いるべきである。R には `lawstat` パッケージに `brunner.munzel.test()` がある。

7 結論と議論

情報教育の枠内で、正しいグラフの描き方や、データから言えること・言えないことを科学的に判断する能力を養う教育ができることを示したつもりである。

実際、情報科の目標のうち、「情報活用の実践力」をデータの正しい視覚化に代表される情報デザイン力、「情報社会に参画する態度」をメディアに溢れるデータを正しく読み解くメディアリテラシー、「情報の科学的な理解」をデータから言えること・言えないことを科学的に判断する能力と理解すれば、本稿で示した内容は

情報科の目標にまさに合致すると言える。

統計的方法といえば、一方では正規分布やそれから導かれる t 分布、 F 分布、 χ^2 分布などを駆使する技術を思い浮かべ、実際のデータは正規分布でないので使えないと誤解されることがある。他方では、多変量データを入力すれば何らかの意味のありそうな結果を出力してくれる因子分析や共分散構造分析などを思い浮かべ、GIGO (Garbage In, Garbage Out) と揶揄されることがある。前者の問題については、中心極限定理を持ち出さなくても、まさにコンピュータの発達によって、分布にかかわらずシミュレーション (並べ替え検定、ブートストラップ) により正確な検定・推定ができる時代になった。これらは、身近な表計算ソフトでも簡単に実行でき、数学的な予備知識もほとんど不要であり、実務はもとより、高校生が統計的方法の核心を理解するのもにも適する。

実際にこのような教育を設計・実施・評価するのは、これからの課題である。

参考文献

- [1] 高木浩光「日常化する NHK の捏造棒グラフ」<http://takagi-hiromitsu.jp/diary/20070128.html>
- [2] 高木浩光「NHK の棒グラフ描画システムが機械的に世論を狂わせている可能性」<http://takagi-hiromitsu.jp/diary/20070204.html>
- [3] 大西俊弘「統計は情報科に移管しよう：PISA 型読解力・メディアリテラシーの向上を目指して」日本科学教育学会年会論文集 Vol. 31, pp. 409–410 (2007)
- [4] 大貫和則「普通教科「情報」における統計リテラシー教育の取り組み」日本科学教育学会年会論文集 Vol. 31, pp. 411–414 (2007)

- [5] 大貫和則「情報科における統計リテラシーを育てるアンケート調査実習」ICT・Education, No. 34, pp. 14–17 (2007)
- [6] 大貫和則「高等学校情報科における統計リテラシー教育」平成 20 年度統計指導者講習会資料, <http://www.stat.go.jp/kids/teacher/pdf/kou.pdf>
- [7] 垣花京子「こんなのでしょうか, 情報の授業—教科「情報」の授業の実態と今後の課題 (2) : 教科「情報」の中での統計教育の導入の検討」日本科学教育学会年会論文集 Vol. 31, pp. 45–46 (2007)
- [8] 成田雅博「教科「情報」に統計教育の視点を取り入れよう」ICT・Education, No. 22, pp. 1–5 (2004)
- [9] 木村捨雄, 村瀬康一郎, 垣花京子, 全国統計教育研究協議会『進む情報化「新しい知の創造」社会の統計リテラシー—教育実践力を高める教師のための統計教育と統計基礎講座』(東洋館出版社, 2005 年)
- [10] Edward R. Tufte, *The Visual Display of Quantitative Information* (Graphics Press, 1983)
- [11] John W. Tukey, *Exploratory Data Analysis* (Addison-Wesley, 1977)
- [12] William S. Cleveland, *The Elements of Graphing Data, 2nd edition* (Hobart Press, 1994)
- [13] Naomi B. Robbins, “Dot Plots: A Useful Alternative to Bar Charts”, <http://www.b-eye-network.com/view/2468> (2006)
- [14] Darrell Huff, *How to Lie with Statistics* (W. W. Norton & Company, 1954); ダレル・ハフ『統計でウソをつく法』(高木秀玄訳, 講談社ブルーバックス, 1968 年)
- [15] 谷岡一郎『「社会調査」のウソ: リサーチ・リテラシーのすすめ』(文春新書, 2000 年)
- [16] Derek Christie, “Resampling with Excel”, *Teaching Statistics*, Vol. 26, No. 1, pp. 9–14 (2004)
- [17] B. D. McCullough, “Microsoft Excel’s ‘Not The Wichmann-Hill’ random number generators,” *Computational Statistics and Data Analysis*, Vol. 52, pp. 4587–4593 (2008)
- [18] R Development Core Team, “R: A Language and Environment for Statistical Computing”, <http://www.R-project.org/>
- [19] 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)