誤用をさけるための注意

権威にしたがうスタイルをやめる

Aという分野の多くの専門家達が問題Xを扱うときに統計的手法Mを使っているという理由で自分もMという手法を使う，というような判断の仕方をしてはいけない．内容を理解してそれが正しい方法であることを論理と証拠に基いて理解できたならば手法Mを使ってもよい．問題のある手法であったことが判明した場合には，その手法を使っている過去の仕事を科学的に批判して，正しい方向に修正する努力をした方がよい．

このノートの内容も「信じて使う」のではなく，まちがっている部分を見付けて修正してから使うようにしてほしい．

P値の誤用を防ぐ

P値を誤用しないためにはASA声明を読むとよい；

The ASA Statement on p-Values: Context, Process, and Purpose
https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

統計的有意性とP値に関するASA声明の日本語訳
https://www.biometrics.gr.jp/news/all/ASA.pdf

ASA声明と疫学研究におけるP値
佐藤俊哉
2017
https://www.jstage.jst.go.jp/article/jjb/38/2/38_109/_pdf

段階検定はできるだけさける（多段階検定での注意）

例 等分散検定 → 2群間のStudentのt検定でP値に補正が必要になる。
直接Welchのt検定を直接的に使う方がよい。

https://biolab.sakura.ne.jp/welch-test.html

https://rstudio-pubs-static.s3.amazonaws.com/57550_7bc5b85e36844a878485b2ad75a4e6f4.html

https://hoxo-m.hatenablog.com/entry/20150217/p1

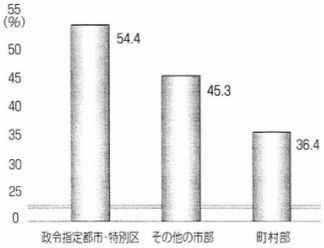
https://oku.edu.mie-u.ac.jp/~okumura/stat/brunner-munzel.html

JuliaによるBurunner-Munzel検定の実装を含む。
https://blog.goo.ne.jp/r-de-r/s/Brunner
https://blog.goo.ne.jp/r-de-r/e/2c2f187d4975cc0928e6f4a0710d6191
https://blog.goo.ne.jp/r-de-r/e/83dc811baf41ecfe469fa794a4c51b84?fm=rss&utm_medium=twitter&utm_source=twitterfeed

http://lbm.ab.a.u-tokyo.ac.jp/~omori/kensyu/nonpara18.htm

不適切なグラフ

不適切な相関係数



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）

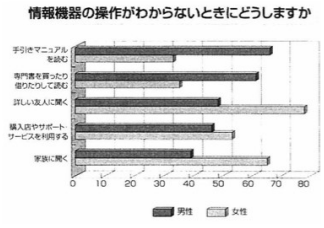


図2　日本文教出版『情報 A』p. 97（2005）。斜めから見ているので目盛りが正確に読めない。横軸の単位がわからない。凡例との対応を色だけに頼っている。

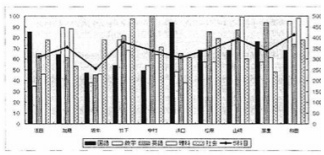


図3　第一学習社『情報 A 整理と実習』p. 108,『情報 C 整理と実習』p. 107（2008）。名義尺度に対して折れ線グラフを使っている。背景が灰色でグラフが見にくい。

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で避けたい。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。

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 $p \fallingdotseq 0.3$ 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 $p \leqq 0.05$ 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 $p = 0.2814$ 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。一様乱数 `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		0.399	6	2		2.167
4	3	4	男	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		0.545	4	5		0.7
9	8	3	女	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

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の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 $p \leqq 0.05$ かだけに意味があるのではない。$p \leqq 0.05$ 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（$0.01 < p \leqq 0.05$ 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 検定を使うかを決める方法は，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）

https://www.google.com/search?q=%E7%B5%B1%E8%A8%88%E6%A4%9C%E5%AE%9A%E3%82%92%E7%90%86%E8%A7%A3%E3%81%9B%E3%81%9A%E3%81%AB%E4%BD%BF%E3%81%A3%E3%81%A6%E3%81%84%E3%82%8B%E4%BA%BA%E3%81%AE%E3%81%9F%E3%82%81%E3%81%AB

統計検定を理解せずに使っている人のために

池田郁男

これは反面教師として非常に参考になる部分がある.

https://twitter.com/genkuroki/status/1155843559129415683

黒木玄 Gen Kuroki @genkuroki · 2019/07/29

統計用語の体系は、専門家の側が寛容になるべき部分が結構あると思います。

例えば「範囲」(range)。←既出

他にも「母数」がparameterの訳語だと知ったときに私は頭を抱えました。

サンプルサイズを「サンプル数」と言ってしまうことも責めたくない感じ。

統計学の本質と全然関係ない話。

1 2 8

リツイート済み

黒木玄 Gen Kuroki
@genkuroki

返信先: @genkurokiさん、@sekibunnteisuuさん、他３人

母集団サイズを「母数」と呼んでいたり、サンプルサイズを「サンプル数」と呼んでいるケースは文脈を読めば、内容を誤解するケースはほぼ皆無だと思います。

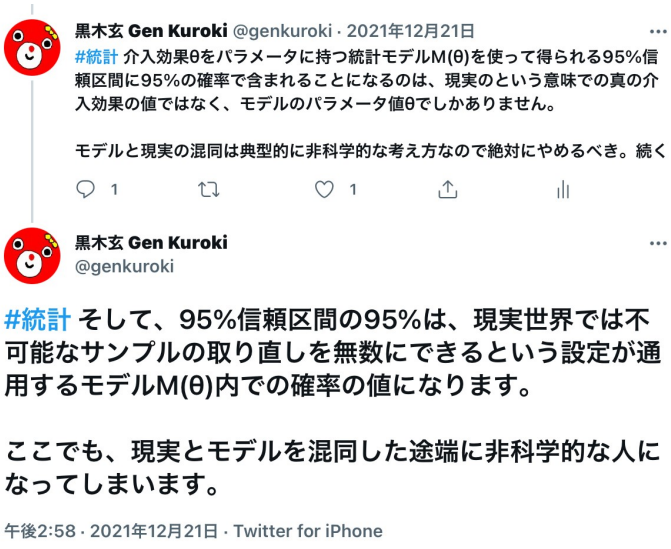
単によろしくない用語の体系に一般人が従っていないだけ。

23:12 · 2019/07/29 · Twitter for iPad

2件のリツイート **4**件のいいね

https://twilog.org/genkuroki/search?word=%E7%8F%BE%E5%AE%9F%20%E3%83%A2%E3%83%87%E3%83%AB%20%E6%B7%B7%E5%90%8C&ao=a

https://twitter.com/genkuroki/status/1473171006877761537



グラフの描き方

- 円グラフはさける.
- legendを見なくて七わかるようにしてるが見易くなる