統計解析環境「R」

中西渉
名古屋高等学校

2011 年 2 月 20 日

1　R とは

R は統計解析に適したプログラミング言語およびその実行環境であり，GPL に基づいて配布されている。S 言語というものを元にして作られたらしい（実物をさわったことがないのでどのようなものかはわからないが）。

2　Why R?

2.1　他の選択肢

統計といえば Excel や SPSS が思い浮かぶことも多いだろうが，高校生に統計を教える場合にそれらが適切な選択肢であるとは思わない。

Excel の統計関数やそのヘルプには多くの問題があったことが指摘されている [1]。今はどの程度改善されたのかわからないが，当時抱いた不信感は未だに根強く残っている。[2] で [3–5] が紹介されていることからすると，今でも Excel は統計用途では使い物にならないのではないだろうか。また [6] にあるような箱ひげ図や幹葉図，ヒストグラムが簡単に描けないし[*1]，一方でひどい 3D グラフが多用される原因を作り出してもいる（特に 3D 円グラフ）。

SPSS は研究などの現場で多く使われているし，高機能であることは間違いない（実物を見たことはないけど）。しかし，とにかく高価だ。大学でも費用の捻出に苦労しているという話を聞くので，高校でそれをまかなうのはかなり難しいだろう。

2.2　R の長所

まず，無料で利用できること，マルチプラットフォーム（Windows，Mac OS X，Linux）であることがあげられる。情報教室はもちろん，教師や生徒が家庭のパソコンにインストールすることについても何の負担もない。

グラフは決して派手ではないが，箱ひげ図やヒストグラムをはじめ，多くの種類のグラフを簡単に作ることができる。また，PDF ファイルや PNG，JPEG 画像として出力すれば文書に取り込むこともできる。

プログラミング言語であることのメリットも大きい。スクリプトを書いておけば，それによって定型的な業務に対応できる。私自身も入試に関する資料作成の一部を R で行なっており，そのスクリプトは年度をまたがって共通で使えるので，過去との比較も容易にできる。

3　デモ

では，実際のデータでどのようなことができるか，実演してみよう。

```
data <- read.csv("pitcher.csv")
mean(data$球速)
boxplot(data$球速)
hist(data$球速)
table(data$球種)
pie(table(data$球種))
by(data$球速,data$球種,mean)
boxplot(球速 ~ 球種,data=data)
```

[*1] Excel2007 の統計アドオンにあるヒストグラムを試したことがあるが，あれをヒストグラムと呼ぶのは. . . 。

4 利用場面

R は学校において，次のような場面で用いることができるだろう。

4.1 統計の授業

新学習指導要領の数学 I「データの分析」では四分位数や箱ひげ図が導入される。R ならそれらを簡単に出力することができる。センター試験（現行）数 II・B の第 5 問でも分布と相関係数の関係がよく出題されるが，たった 10 個の資料で相関というのも無理がある。センター試験では手計算できなくてはいけないから 10 個なのだが，パソコンを使わないと処理しきれないような数の資料を扱ってこそ統計の意味がある。

4.2 シミュレーション

プログラミング環境であり，ベクトルを柔軟に取り扱うことができるので簡単に数理モデルのシミュレーションを行なうことができる。たとえば，ランダムに打った点が円の中に入る割合から円周率の近似値を求めるシミュレーションは

```
x <- runif(1000)
y <- runif(1000)
mean(x^2 + y^2 < 1) * 4
```

というようにループも使わずに書ける [2]。

以前，表計算のマクロなどを用いて統計の実習教材を作ったことがある [7]。しかしマクロのコードを理解させることはハードルが高いので，生徒にしてみればブラックボックスであるプログラムを操作しているだけだった。しかし R なら少しのコマンドを並べるだけでそういった「実験」ができる。たとえば標本平均の様子を調べたいなら，次のような計算をしてみればいい [8]。これは平均 50，標準偏差 10 の母集団から取り出した大きさ 100 の標本の平均を 10000 回とって，その標準偏差を求めるものである[*2]。

```
x <- numeric(length=10000)
for(i in 1:10000){
  x[i]<-
    mean(rnorm(n=100,mean=50,sd=10))
}
mean((x-mean(x))^2)
```

4.3 業務

我々教員の日常業務において，統計処理を必要とする場面はいくらでもある。まずは自分の手元の作業を R でやってみてはどうか。自分がやらないこと，必要性を感じないことを生徒に教えても何の意味もないのだから。

5 情報入手先

5.1 Web

R の公式サイト [9] にドキュメントがある。日本語のサイトで参考になるものも多いので，あちこち拾い読みするのがいいと思われる [10–13]。

5.2 書籍

紙の本が手元にあると何かと安心である。私は [8, 14] で勉強しているが，以前に比べればたくさんの本が出版されている。R を中心にした本もあるし，統計学を中心にした本もあるので，自分に合ったものを選ぶのがいい。

5.3 勉強会

名古屋大学大学院国際開発研究科の阪上辰也氏が中心になって Nagoya.R という勉強会が数回催された。毎回前半に初心者講習会を行なっていて，そのテキストは [15] で公開されている。これを見て実際に操作をすることで，基本中の基本は押さえることができるだろう。

Nagoya.R 以外にも Tsukuba.R，Tokyo.R，Osaka.R などの地方コミュニティがあり，2010 年 11 月にはそれらが集まって Japan.R という勉強会が開かれた。

6 短所

R の導入を検討するにあたって，現時点で問題になりそうなことをいくつかあげておく。

[*2] 標準偏差の計算に `sd` を用いていない理由は後述。

■CUI　設定をすべて関数やコマンドに引数として与えなくてはいけないことが，ハードルの高さになってしまう。R Commander や Excel アドインの RExcel を用いる方法もあるのかもしれないが，それはそれで別の面倒を生むかもしれない。

■日本語対応　R では変数にも日本語が使えるし，日本語を含むデータを扱うこともできる。ただし文字列処理に関してはいくらか疑問に思うところがないこともない [16]。また，ヘルプが英語であることも一つの壁となるだろう。

それ以外にも次のような問題があるので注意されたい（Windows 特有の現象かもしれない）。

- Windows 用インストーラを日本語で実行すると途中で文字化けする（読めなくてもそのまま進めて問題ないのだが，気持ち悪いので言語を聞かれたときに 'English' を選択する。それでも実行環境は日本語になるので問題ない）。
- GUI 設定でフォントを「MS Mincho」など日本語文字を含むものに変更しないと，いくつかの場面で文字化けする。

■資料　発行される書籍がかなり増えたとはいえ，手頃な内容のものが多くあるとはいいがたい。大学生以上を対象とした本ばかりで，高校生に適切なものが見つけられないことが不満だ。いいものがあったら是非とも教えてほしい。

■質問できる人　困ったときに質問できる人が職場内にいるかというと，正直厳しいと言わざるを得ない（ネットを通じてなら可能だが）。実際には，今の時点で使い始める人がむしろ質問される側の人になるのだろう。

■分散・標準偏差　`var` や `sd` で計算される分散や標準偏差は，分母が n ではなく $n-1$ である。このことは欠点ではないのだが，高校の教科書では分母が n のものしか扱わないのだから，その点を補う必要はある。具体的には

```
varp = function(x) {
  var(x)*(length(x)-1)/length(x) }
stdevp = function(x) {
  sqrt(varp(x)) }
```

などの形で定義しておけばいいだろう [2]。

7　まとめ

現時点では新指導要領数学 I の「データの分析」がどのような扱いになるのか正確にはわからない。漏れ聞くところによると，この単元をできるだけ軽く扱えるように**配慮**している教科書が作られているという。なるほど，それは**一部の教師**に需要がありそうだ。同僚に話を聞いても，四分位数など新しい内容[*3]についてよく知っている者は少ない。新しい内容を勉強したくない教師にとっては，申し訳程度の内容で済ませられる教科書がありがたいのだろう。

しかし，日本の教育で今まで余りにも統計をやらなさすぎたことが問題なのだ。それは今に始まったことではなく，私の世代でも「確率・統計」は $\sum$ の計算の応用にすぎず，推定・検定は授業で扱われなかった。そのためか，マスコミにはデタラメなグラフがあふれている。[17] は 40 年も前から読まれているというのに。このことについて，情報教育に携わる者の何人かは警鐘を鳴らし続けているし，私も機会を見つけては意見を述べるようにしてきた [7, 18, 19]。

我々の仕事は生徒が「知るべきこと」を教えることであるのに，どうも「知ってること」に偏りがちであるように思われるのだ。統計ソフトといったときに Excel と短絡してしまうのもその一つではないか。

学びをやめた者は教壇を降りなくてはいけない。

[*3] といっても 30 年以上前のものだが

参考文献

[1] 青木繁伸「Microsoft Excel を使った統計解析」
http://aoki2,si.gunma-u.ac.jp/lecture/stats-by-excel/

[2] 奥村晴彦「R を使った情報教育」，情報処理学会「SSS2010 情報教育シンポジウム論文集」pp.77–80, ISSN 1344–0640

[3] B.D. McCullough and David A. Heiser, "On the accuracy of statical procedures in Microsoft Excel 2007," *Computational Statistics and Data Analysys* **52**, 4570–4578(2008).

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

[5] B.D. McCullough "Microsoft Excel's 'Not The Wichmann-Hill' random number generators," *Computational Statistics and Data Analysys* **52**, 4587–4593(2008).

[6] John Wilder Tukey, "Exploratory Data Analysis", Addison Wesley, 1977.

[7] 中西渉「情報の授業で統計を扱う」，情報処理学会「SSS2007 情報教育シンポジウム論文集」pp.11–14, ISSN 1344–0640

[8] 山田剛史，杉沢武俊，村井潤一郎『R によるやさしい統計学』（オーム社，2008）

[9] R Development Core Team, *The R Project for Statistical Computing.* http://www.r-project.org

[10] 奥村晴彦「統計・データ解析」http://oku.edu.mie-u.ac.jp/~okumura/stat/

[11] 青木繁伸「R による統計処理」http://aoki2.si.gunma-u.ac.jp/R/

[12] 舟尾暢男，*R–Tips* http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

[13] 岡田昌史，*RjpWiki* http://www.okada.jp.org/RWiki/

[14] 舟尾暢男，高浪洋平『データ解析環境「R」』（工学社，2005）

[15] *Nagoya.R Wiki* http://corpus-study.info/nagoyar/

[16] 安部晃生「R で 2 ちゃんねるを読んでみた」http://blog.recyclebin.jp/archives/1375

[17] ダレル・ハフ『統計でウソをつく法』（講談社ブルーバックス，1968）

[18] 中西渉「情報の授業における統計の扱い」http://www.meigaku.ac.jp/~watayan/doc/20061021/ 2006 年度・第 55 次愛知県教育研究集会

[19] 中西渉「こまけぇこたぁいいんだよ」http://www.meigaku.ac.jp/~watayan/doc/20100508/ Nagoya.R #2 (2010)