統計言語 R の心理統計での利用について

岩 橋 俊 哉

The statistical language "R" and psychological statistics.

Toshiya IWAHASHI

I．統計言語 R の概要

最近、統計言語 R が注目されるようになってきた。その理由は、R が著名な統計言語のひとつである S の互換機能を持っていながら、オープンソースの手法で開発されており、GNU の GPL のもとに配付されていることにある。この点で学校など学術機関での利用には最適であると考えられる。なぜなら、（正確には単純に無料というわけではないのだが、）無料で使う事もでき、そのソースも公開されているので、解析、改良などが可能であること。また、S や商用版の S-plus とは異なり、複数の OS（Unix, Windows, MacOS）で開発が進められているので、OS の制約なく、広く利用できることも大きな長所である。このことから、特に学生などが高価な統計用アプリケーションを購入しなくとも好きな機器で簡単に高度な統計計算を試すことができるようになった。参考までに、心理学でよく使われている代表的な統計パッケージの価格を書いておくと、SPSS は、教育機関での価格で、最低 10 万円程度から数十万円である。S の商用版である S-plus も同じ価格帯であり、SAS については、機関を対象としたリースであり、個人を対象としていない。その他の統計パッケージの場合でも 10 万円程度の価格の物が多く、さらにバージョンアップ時の費用などを考えると、特に個人では気軽に購入しにくいのが現状である。

S は、John M. Chambers, と Becker により、1980 年前半に開発されたデータ解析環境である。SPSS や SAS のように当初は、大型コンピュータ用に開発されたものではなく、最初から Unix ワークステーションでの操作を想定して開発されており、インタラクティブに操作できることが特徴のひとつである。

R は、当初ニュージーランドのオークランド大学統計学部門の Ross Ihaka と Robert Gentlemanat により S の互換環境として開発されたものである。その後、多くのボランティアによってコードの吟味やテストがなされ、1997 年後半からは、CRAN（http://cran.au.r-project.org/）が R のコードを管理、配付している。

Rは（Sと同様に）、文法が比較的単純な上に、拡張性があり、ユーザーは新しい関数を定義して機能を追加できるようになっている。システムの大部分は、Rそのもので書かれており、ユーザーは使われているアルゴリズムを容易に確認でき、さらに他者とアルゴリズムを共有できるというメリットもある。

問題点としては、オープンソース故に動作に保証がない点だろう。特に、数値の精度など結果の信頼性に関わる点においては、企業や公的機関などで利用する場合には問題になるだろう。その他の場合にも、どの程度の大きさのデータまで処理することができるかなど、明確ではない点がある。ただ、この点はRの普及が進むにつれて克服されていく問題ではあるが。

II. 教科書案

心理学の初学者が統計を学ぶツールとして、Rがどの程度使えるか、教科書の案を作成してみた。現在では、ExcelsとSPSSが統計の初学者用のツールとしてよく用いられているものと思われる。Excelの長所は、その直感的な操作性にある。また、広く普及しているアプリケーションであることから、参考書などが豊富であるという利点もある。ただし、統計処理の機能が不足しているため、SPSSで補って、統計処理を行なうということが現在ではよく行われている。SPSSはExcelの操作性をとり入れて、簡単に統計処理ができるように設計されている。さらにExcelのファイルを変換せずに読み込めるという特長もある。それに比べると、Rの操作性はやや敷居が高い感がある。その代わり、費用をかけずに、単独で、高度な統計処理まで行えるというのはそれを充分に補うメリットである。

1. Rの用語と操作方法

ここでは、現時点でのRの最新版2.2.1（Macintosh版）を対象としている。

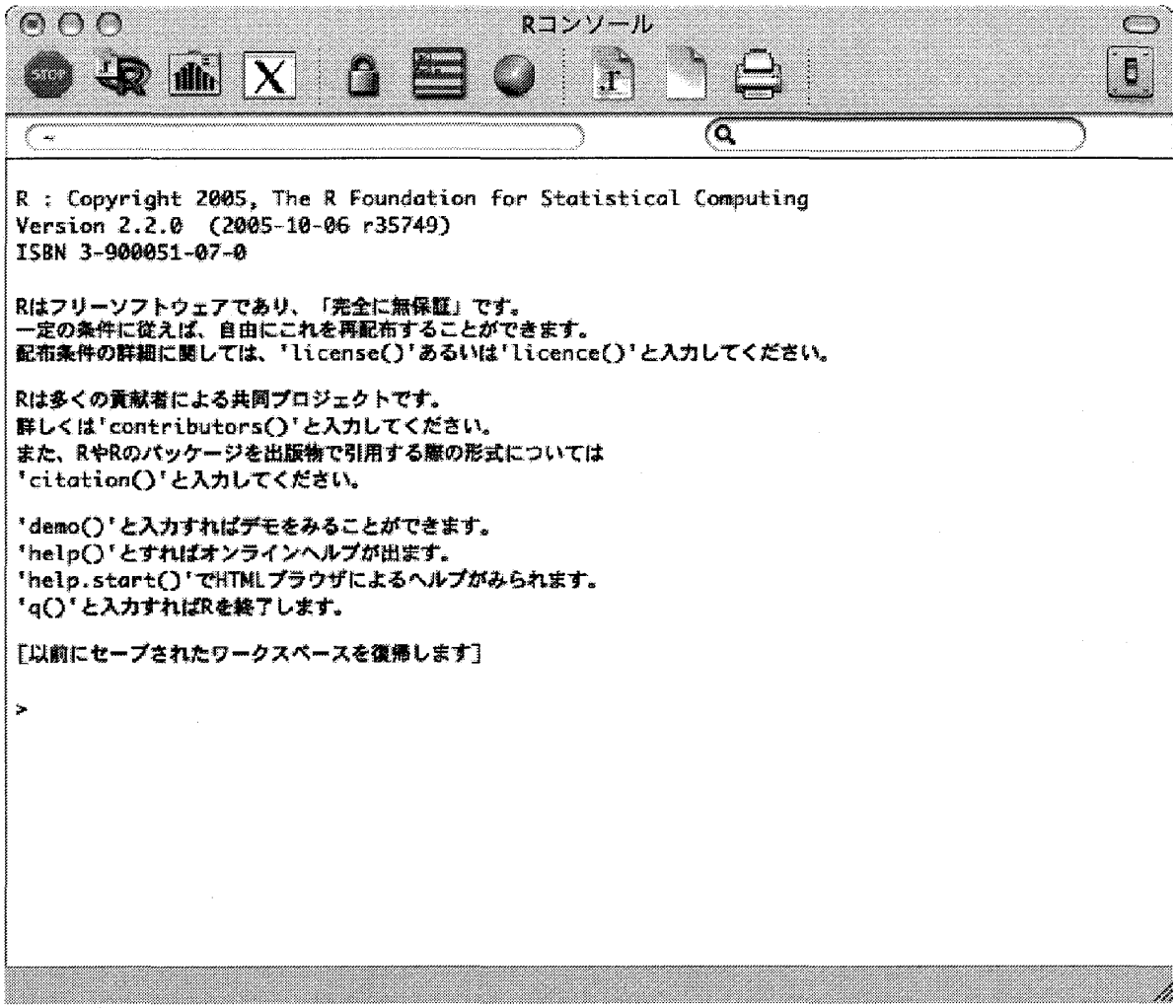
起動方法　アイコンをダブルクリックする。

終了方法

```
>q()
```

と入力する。

画面の構成は、ほぼOSによらず共通である。起動すると以下の図のような、Rコンソールが表示されるので、そのプロンプト（>）の後にコマンドを入力する。直接キィボードから入力して、電卓的に用いることも可能。ただ、長いスクリプトの場合には、エディターで編集してから、実行する。



2. 文法の主な特徴

・関数の使用方法を覚えるだけで大抵の統計処理ができるようになっている。

・代入の記号には通常、等号（=）ではなく、 次の記号（<- ） を用いる。

・ベクトル変数の扱いが容易である。

・データ・フレームがExcelのワークシートにあたると考えてよい。

・変数名などに日本語は、まだ使いにくいので、原則避けた方が良い。

3. 実習

A. コマンドの直接入力

以下のコマンドをキィボードから入力する。この使い方は、Excelに比べて手軽に入力できる点が優位かもしれない。この入力方法には、ヒストリー機能があるので、矢印キィで文字列を呼び出して簡単に入力文字を編集することができる。

例、最初に以下のコマンド（下線部分）を入力する。

> a <- c (1,2,3,4,5))

これを入力すると、変数 a に　数値 1,2,3,4,5 が代入される。c は、配列を設定する関数である。

次に、代表値の関数を入力してみる。平均を求める場合は、

>mean (a)

最大値を求める場合は

>max (a)

最小値を求める場合は

>min (a)

とそれぞれ入力することで結果が得られる。

また、

>summary (a)

と入力すると、最小値、第一四分位数、中央値、第三四分位数、平均値が一度に算出される。

ちなみに Excel の「データ分析」の基本統計量では、平均 , 標準誤差 , 中央値 (メジアン) , 最頻値 (モード) , 標準偏差 , 分散 , 尖度 , 歪度 , 範囲、最大、最小、合計、標本数が算出される。

B. エディターを利用した入力と実行

R のプログラムをあらかじめテキストファイルで作成し、保存したデータを「ソースを読み込む」で読み込んでから、「編集」メニューの「実行」で実行する。

C. データを読み込んで、処理する。

データ例、Excel で以下のように表記できるデータを処理する。

ID	性別	国語	算数
1	F	55	70
2	M	60	85
3	M	80	60
4	M	65	75
5	F	50	80

R では、この書式のデータは、データフレームと呼ぶ。作成方法は、

a. 以下のようにベクトル変数を組み合わせて、直接データフレームで作成する方法と

例、

> sex <- c (“F”,”M”,”M”,”M”,”F”)

> kokugo <- c (55,60,80,65,50)

> sansuu <- c (70,85,60,75,80)

> Seiseki <- data.frame (Kokugo=kokugo, Sansuu=sansuu)

b. テキストファイルを読み込んで作成する方法がある。

例、あらかじめ、エディターで、csv 書式のデータ (seiseki.csv) を作成しておく。

ID, Sex, kokugo, sansuu

1, “F”, 55, 70

2, “M”, 60, 85

3, “M”, 80, 60

4, “M”, 65, 75

5, “F”, 50, 80

> read.csv (“seiseki.csv”)

データファイルの書式については csv 書式などのテキストファイルが良いと思われる。R には拡張機能があるので、Excel のファイルを直接読むように設定することはできるが、他のアプリケーションとデータを共有することを考えると、ほとんどのアプリケーションで標準で読み書きできるテキストファイルの利用が望ましい。

D. データ・フレームのデータを処理する

変数による処理の指定　結果の出力。入力すると以下のように要約統計量が表示される。

> summary (Seiseki)

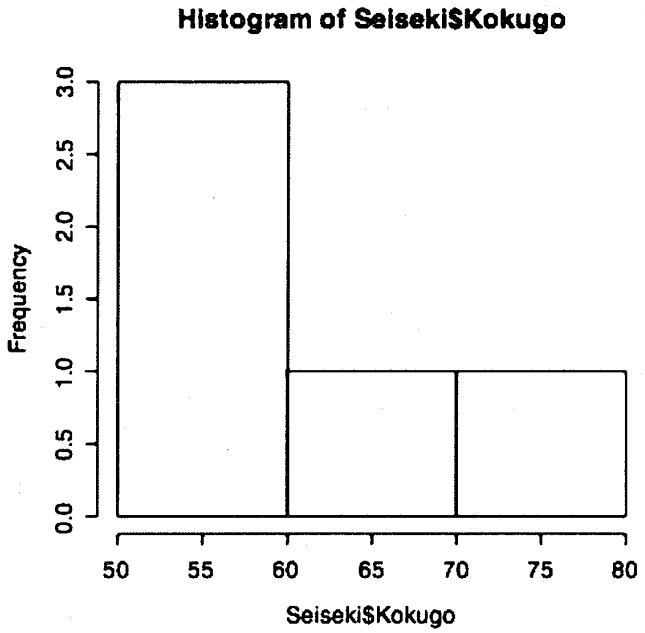
Sex	Kokugo		Sansuu	
F:2	Min.	:50	Min.	:60
M:3	1st Qu.	:55	1st Qu.	:70
	Median	:60	Median	:75
	Mean	:62	Mean	:74
	3rd Qu.	:65	3rd Qu.	:80
	Max.	:80	Max.	:85

E. グラフの作成方法

ヒストグラムを作成する場合には、

> hist (Seiseki$Kokugo)

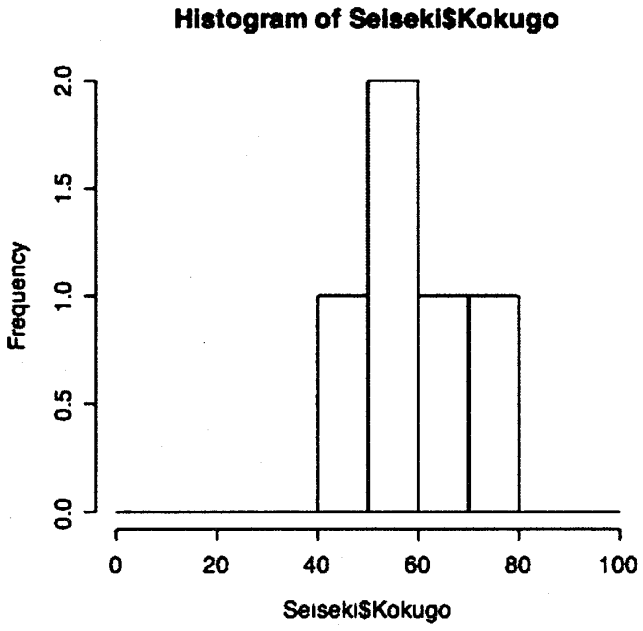
と入力すると、グラフ用のウィンドウに以下のようなグラフが表示される。



さらに区間などを指定する場合は、以下のように指定する。

> hist (Seiseki$Kokugo, breaks = seq (0, 100, by = 10)

上記の seq 以後のように区間の最大値、最小値、幅を指定することができる。



4. その他の主なグラフ描画関数として、以下の２種類のものがある。

A. standard plots（グラフそのものを描画する関数）

plot（）散布図など、hist（） 度数分布、boxplot（）箱ひげ図、barplot（） 棒グラフ、piechart（）円グラフ、plotmeans（） 平均と標準偏差を表示する。ただし、この関数はライブラリを追加する必要がある。

B. plotting elements（グラフの中に直線、四角形などの図形を表示する関数）

lines（）直線、points（） 点、arrows（） 矢印、title（） グラフのタイトルなど

5. 主な記述統計用の関数として、以下の関数がある。

mean（）平均、sd（）標準偏差、var（） 分散、median（） 中央値、quantile（） 四分位偏差、cor（） 相関係数（Spearman, Kendall の順位相関係数も含む）、table（） クロス集計

6. 主な検定用の関数（心理学でよく用いられるものに限っている）

t.test（） t 検定

pairwise.t.test（） 対応のある t 検定

cor.test（） 相関係数の検定

var.test（） f 検定

lm（y ~ x） 回帰分析

lm（y ~ f） 一元配置の分散分析

lm（y ~ f1 + f2） 二元配置の分散分析

lm（y ~ f + x） 共分散分析

主な変量解析

lm（y ~ x1 + x2 + x3）重回帰分析

princomp（）, prcomp（） 因子分析、主成分分析

factanal（） 最尤法により因子を求めプロマックス回転する場合。

ノンパラメトリック検定

wilcox.test（） ウィルコクソンのサイン・ランク検定

kruscal.test（） クラスカル・ワリスの検定

friedman.test（） フリードマンの検定

III. 最後に

初学者が統計を学ぶツールとして、R は対費用効果、その機能、使い勝手の点で、他の統計アプリケーションと比べて遜色ないものと思われる。問題は、現時点での普及率の低さであるが、サイトや書籍などの情報源は、徐々にだが、充実してきており、R の特長を考えると、今後さらに普及が進むものと思われる。今後の課題は、特定の分野に合わせた教科書や解説書が出版されることである。

参考文献及び URL

文献

Dalgaard, P., 2002, *Introductory Statistics with R*, Springer.

舟尾 暢男・高浪洋平 2005 「データ解析環境「R」—定番フリーソフトの基本操作からグラフィックス、統計解析まで」工学社

岡田 昌史　2004 「The R Book—データ解析環境 R の活用事例集」　九天社

竹内 俊彦　2005「はじめての S-PLUS/R 言語プログラミング—例題で学ぶ S-PLUS/R 言語の基本」オーム社

URL

http://cran.au.r-project.org/

The Comprehensive R Archive Network R 言語の公式サイト

http://www.okada.jp.org/RWiki/index.php?RjpWiki

日本語による R の情報交換サイト

http://aoki2.si.gunma-u.ac.jp/

群馬大学　青木繁伸氏のサイト特に Mac での R の利用などに詳しい。「統計学自習ノート」の中の「R function」を参照

http://mat.isc.chubu.ac.jp/R/tech.html

松井孝雄氏　R による分散分析例が載っている。

http://cat.zero.ad.jp/~zak52549/R.html

Okamura Yasuyuki 氏 「無料統計ソフト R（CRAN）で心理学 Passepied」

http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

竹澤邦夫氏 「R の備忘録頁」

http://cwoweb2.bai.ne.jp/~jgb11101/index2.html

前述舟尾暢男氏のサイト

（2006 年 9 月 25 日受理）