R 言語

則行祐作，内村祐之
Noriyuki Yusaku，Uchimura Yushi

1　はじめに

統計解析にコンピュータが利用され始めた頃，多くの研究者は C 言語といった汎用的なプログラミング言語で解析を行った．しかし C 言語で統計解析のプログラムを記述すると，膨大な記述量になり効率が悪い．さらに C 言語はその言語について一通り学ぶ必要があり，プログラミングに素養がない人には困難であった．このような背景から，誰でも容易に扱える統計解析言語の R 言語が開発された．

2　R 言語とは

2.1　概要

1990 年代にニュージーランドのオークランド大学の Ross Ihaka とアメリカのハーバード大学の Robert Gentleman によって開発された．R 言語は統計解析に特化した言語である．データを効率的に操作・保管するための仕組みや，配列や行列の演算をサポートした演算子のセット，結果を可視化するためのグラフ作成機能などを備えている．プログラミング言語としては条件分岐やループ処理，ユーザー定義の再帰的関数，テキスト形式のデータ入出力などの基本的な機能を備えているほか，オブジェクト指向も取り入れている。また多くの統計手法を実現する関数を標準でサポートしており，簡易なコマンドで実行可能である．現在，R 言語は多くの統計解析に用いられている．具体的な利用例としてデータマイニングや機械学習が挙げられる．

R 言語の主な特徴を以下に 3 つ挙げる．

- オープンソース
- 多様な関数
- CRAN（The Comprehensive R Archive Network）

以上 3 つの特徴は R 言語の普及を支えている特徴であり，今後の普及にも繋がると考えられる．

2.2　オープンソース

R 言語はオープンソースフリーウェアであり，無料で利用できる．統計解析言語・システムは，R 言語の他に SAS（Statistical Analysis System），SPSS（Statistical Package for Social Science）が挙げられる．しかし，この 2 つの統計解析言語・システムは高価である．例えば学生が上記のような高価なソフトウェアを用い大学等で実習を受けたとする．しかし卒業などで環境が変化すると，習得した技術を活用することが多くの場合困難になる．一方，R 言語は，無料であることで，ユーザの技術が無駄にならず，どんな環境でも利用することができる．

次にオープンソースによる利点として，3 つ挙げる．1 つはソースコードが公開されているので，全世界の開発者によってバグが発見され，修正されることを可能にする．すなわち R 言語は信頼性が高い．2 つめは R 言語を他言語で利用するためのライブラリが作成されることである．例として Ruby の RSRuby，Python の rpy2 などがある．R 言語のライブラリを利用するプログラミング言語は，いずれも汎用的である．そのため統計処理以外の処理を元言語で行いながら，R 言語が得意とする統計処理のみ任せることが可能である．3 つめに無料であることから，利用者が多く参考書が豊富である．以下の Table1 に近年統計解析に使われる言語の参考書数を示す．

Table1　出版日が 2010 年度以降の参考書数

言語	参考書数
SPSS	750
SAS	350
Python	20
R	3700

この Table1 は”Amazon.com”[1)] の”books”のカテゴリで「”言語名” statistics」と検索した結果である．この表を見ると，R 言語の参考書数が飛び抜けて多いことが分かる．これはプログラムが分からないときに参考にできる書籍が多く，疑問の解答を得やすいといえる．プログラム外の部分で支持される理由である．

2.3　多様な関数

次に標準で実装されている多様な用途の関数の一例を Table2 に示す．

Table2　R の関数表

関数名	目的
mean()	平均を求める
var()	分散を求める
sd()	標準偏差を求める
plot()	グラフを描画
t.test()	二標本 t 検定
cor.test()	無相関検定
chisq.test()	適合度検定

Table2 に示すように，統計解析を簡潔に行える関数が本体に組み込まれている．他の言語では基礎的な平方根，累乗計算などの基礎的な数値計算関数を組合せて，統計

計算を行う．しかし R 言語は，この Table2 のような関数に数値や統計データを入れることで数値計算や統計検定を可能にする．さらに標準に関数が組み込まれているため，統計処理以外のコードを書く必要がない．また統計検定の式を知らなかったとしても，方法を知っているだけで検定を行うことが可能である．

2.4 CRAN[2)]

最後に CRAN について述べる．CRAN とは，R 言語のパッケージや異なる OS 用のバイナリデータを提供するネットワークである．統計解析手法を実現するために開発したパッケージが公開されており，個人で開発した関数を CRAN にアップロードできる．他の言語の場合，個人の作者が作成したパッケージを利用するとき，複数の Web サイトで目的のパッケージを見つけてダウンロードする必要がある．しかし CRAN に全ての R 言語のパッケージが存在する．また日々更新されており，統計解析手法の進歩に応じた関数が即利用できる．すなわち自分の環境にダウンロードすることで，最先端の R の統計関数を使用することが可能である．

3 プログラムの実行例

2 章で挙げた関数を実際に実行したプログラムを以下の Table3 に示す．

Table3 R のソースコード

```
 > x <- read.csv("C:/Program Files/R/data.csv")
 > plot(x,
 +      xlab="3 月の日付 [日]",
 +      ylab="気温 [℃]",
 +      ylim=c(1,20),
 +      type="l",
 +      axes=F)
> axis(1, pos = 0, at =0:31, tcl=0.5)
> axis(2, pos = 0, at =0:18, tcl=0.5)
> mean(x[,2])
[1] 8.383871
> var(x[,2])
[1] 12.8374
> sd(x[,2])
[1] 3.582931
```

この Table3 のプログラムは線グラフに描画し統計処理を行うものである．京田辺市の 2014 年 3 月の平均気温のデータを気象庁のホームページからダウンロードした．以下の Fig.1 に作成したグラフを示す．

下記に Table3 で示したプログラム内容について述べる．このプログラムでは，read.csv 関数，plot 関数，axis 関数，mean 関数，var 関数，sd 関数を用いている．これらの関数を用いることでグラフの描画と統計解析の実現を可能にしている．まず，今回，read.csv 関数でデータを読み込んだ．plot（）関数には，x 軸，y 軸のラベル，

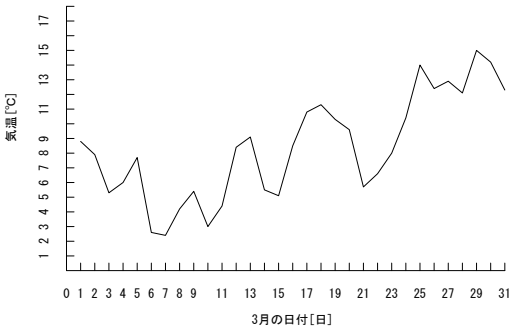

Fig.1 R で表示したグラフ

y 軸の範囲，そして線グラフの 4 つの項目を設定した．2 つの axis（）関数では，それぞれ x 軸，y 軸のメモリの開始位置，メモリの範囲，内向きメモリの長さを設定した．

最後に，mean()，var()，sd() 関数を用い，このデータに統計処理をおこなった．mean() は平均，var() は分散，sd() 標準偏差を求める関数である．平均は，全てのデータをデータ数で割った値である．また分散と標準偏差はデータがどれだけばらついているかを表しており，平均±標準偏差の二倍の範囲に約 95 パーセントのデータが分布している．

4 展望とまとめ

データを解析し我々の生活に利用するために，日々新たな統計解析手法が生まれている．R 言語は，オープンソース，多様な統計関数，CRAN という 3 つの特徴を持つ．これらの特徴により，この進歩に対応していく．例えば，新しい統計解析手法に対応した関数が含まれたパッケージをユーザが CRAN にアップロードする．アップロードした中で有益な関数が多く使われることで，R 言語のアップデートにより標準の関数に実装される．さらに他の言語にライブラリとして追加されることで，幅広いユーザに使われる．このように R 言語は誰にでも容易に扱える統計解析言語としてだけでなく，他の言語にも利用される．そして言語の枠を超え，統計解析の手段の中心としてこれからも発展していくであろう．

参考文献

1) Amazon.com
   http://www.amazon.com/
2) CRAN
   http://CRAN.R-project.org/
3) 藤井良宜，金明哲 『カテゴリカル・データ解析 -R で学ぶデータサイエンス-』（共立出版，2010）
4) ウーヴェ・リゲス 『R の基礎とプログラミング技法』（シュプリンガー・ジャパン，2006）