# レポート課題 1 「統計解析環境 R の使い方」

平井有三

## 1 何故 R

1. 無償ソフトウエア

   http://cran.r-project.org/ ::ソフトウエアの総本山
   http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html ::充実したマニュアル
   統合開発環境 RStudio が使える。便利なので R と共にインストールしましょう。下記のリンクに分かりやすい初期設定や使い方の説明がありますので、是非参考にしてください。

   https://das-kino.hatenablog.com/entry/2019/11/07/125044

   なお、代入演算子 <- のショートカットは、Alt + - です。+ は同時に押すという意味で、+ を入力する必要はありません。

2. インタープリタ型言語： 暴走しない
3. プログラムが容易

   >c(1:10) >は r のコンソール入力のプロンプト。c() はベクトルを作るコマンド
   [1] 1 2 3 4 5 6 7 8 9 10 コマンドの実行結果
   >2^c(1:10) ベクトルの要素ごとに作用させる
   [1] 2 4 8 16 32 64 128 256 512 1024
   >c(1:10)[c(2:5)] [ ] は配列要素の指定、ベクトルの一部を抽出する
   [1] 2 3 4 5

4. 統計解析から出発。MATLAB は制御が出発点。
5. ベクトル、行列演算が得意

   >a<-seq(1,10,2) :: 代入は <- または ->または=。伝統的には<-の使用が、最近は=の使用が推奨されているようである。

```
>a
[1] 1 3 5 7 9
>b<-seq(-1,,2,5)
>b
[1] -1 1 3 5 7
>a+b
[1] 0 4 8 12 16
>a*b
[1] -1 3 15 35 63
>a/b
[1] -1.000000 3.000000 1.666667 1.400000 1.285714
>a^b
[1] 1 3 125 16807 4782969
>a %*% b ::ベクトル a と b の内積。結果は 1 行 1 列の行列
     [,1]
[1,]  115
```

6. 統計ライブラリーが豊富

```
>library()
>library(rgl)
```

7. グラフィックスのサポートが充実しており、統計関係のグラフが簡単に作成できる
8. 標準乱数発生器がすごい (Mersenne-Twister)
   - 松本眞、西村拓士による発明
   - 周期 : $2^{19937} - 1 \simeq 4.31 \times 10^{6001}$
   - 1 億 $(10^8)$ の日本人が毎秒 1 兆個 $(10^{12})$ の乱数を 100 年 $(3.65 \times 10^4)$ 使い続けても、必要な乱数の数は $3.65 \times 10^{24}$ を超えない

# 2 基本操作

## 2.1 起動と終了

1. 起動 :
   - Windows: R アイコンのクリック
   - MacOS: アプリケーションフォルダーで R.app を選択
2. 終了
   - コンソールで ＞ q()

- 現在の作業空間内のオブジェクトを現在の作業ディレクトリーに保存するか否か聞かれる。保存すれば、次回オブジェクトとコマンド履歴を復元することが可能である。

3. ヘルプ：

```
>help(関数名)
>library(help=ライブラリ名)
>data()  ::利用可能なデータセットを表示
```

## 2.2 ディレクトリ関連

以下のコンソールコマンドには、RGui のメニューバーから利用可能なものがたくさんあるので、各自調べておくこと。

1. 現在の作業ディレクトリの表示：

```
>getwd()
```

2. ディレクトリの設定：

```
>setwd(‘‘ディレクトリの絶対・相対パス’’)
```

3. 現在の作業ディレクトリ内のファイルの表示：

```
>list.files()
```

4. コンソールから入力したコマンド履歴の保存と読み込み
   コンソールから入力したコマンドは、上矢印キーで一つ前のコマンドを、下矢印キーで一つ後のコマンドを表示することができる。必要なコマンドを表示して編集し、リターンキーで再実行すれば、効率的な計算を行うことができる。
   - 現在のディレクトリーの filename.Rhistory に保存：
     `>savehistory(file="filename.Rhistory")`
   - 現在のディレクトリーの filename.Rhistory から読み込み：
     `>loadhistory(file="filename.Rhistory")`
     読み込んだコマンド履歴は、矢印キーなどで必要なコマンドを表示し編集・再実行できる。大切な手順などをファイル名を変えて保存して利用するとよい。
   - コマンド履歴の表示：
     `>history()`
5. 起動から終了までのワークスペースの明示的保存と読み込み
   - 保存：
     `>save.image(‘‘intro.RData’’)`

- 読み込み：
  `>load(‘‘intro.RData’’)`
- ワークスペース内のオブジェクトの確認：
  `>ls()`
- ワークスペース内オブジェクトの消去：
  `>remove(‘‘オブジェクト名’’)`
  `>remove(list=ls())  ::全オブジェクトの消去`

6. エディタで作成した関数 (識別子として.r を用いる) の読み込み：

```
>source(‘‘ファイル名.r’’)
```

## 2.3 ライブラリーのインストール法

新しいライブラリーを、自分のホームディレクトリーの $HOME/R/Library にインストールしたい場合の方法。詳しくは help を参照のこと。

1. R の環境設定で、起動の項目中の、「デフォルトのライブラリーパス」で「$HOME/R/Library を追加」にチェックする。
2. メニューバーの「パッケージとデータ」から「パッケージインストーラ」を選択すると、「R パッケージインストーラ」ダイアログが表示される。
3.「一覧を取得」ボタンを押す。
4. CRAN のミラーサイトを聞かれたら、「japan(Tokyo)」を選択する。以前は、筑波大学医学部にもミラーサイトがあった。
5. 表示されたリストから必要なパッケージを選択する。左ボタンのクリックで選択されたライブラリーの行が青く変わる。他の行で左ボタンをクリックするとその行のライブラリーが選択され、前に選択されたものは解除される。複数のライブラリーを選択したい場合は、コマンドキーを押しながら左ボタンをクリックする。連続した行のライブラリーを選択する場合は、shift キーを押しながら左クリックすればよい。
6. e1071 ライブラリーが選択されたものとする。
7.「インストールする場所」を「ユーザエリア」にする。
8.「選択をインストール」ボタンを押すと、「デフォルトのライブラリーパス」内にインストールされる。
9. メニューバーの「パッケージとデータ」から「パッケージマネージャ」を選択し、一覧の更新ボタンを押す。
10. 状態が「未ロード」の場合、左のチェックボックスをチェックすると、ライブラリーがロードされた状態になり、R コンソールから使えるようになる。R コンソールから library() コマンドでロードしてもよい。

# 3　データフレームと散布図の表示

データフレームは各行・各列にラベル（名前）を付けられた 2 次元配列（表）である。各行が一つのベクトルデータを、各列がそのデータの各ベクトル要素（属性）の値（属性値）を表す。これらのラベルを用いて、指定した行や列の抽出など、指定されたデータに対する一括操作を行うことができる。R はデータフレームで表現された統計データに対して種々の統計処理を行うことができるシステムである。

本節では、データフレームから属性対ごとにデータの散布図を表示し、データ間の関係を視覚化する手法について、課題を解きながら修得する。

本講義では、ここで説明に使用するアヤメデータセットを種々の統計処理の例題として使用する予定である。他の利用可能なデータセットも

```
>data()
```

コマンドを実行すれば確認できるので、興味のあるデータについて統計処理を行ってみること。また、新しいライブラリーをインストールすると、新しいデータフレームが付いてくるので、そちらも参考にすること。

データフレームの内容は、relimp というパッケージの中の関数 showData() を用いると、別ウィンドウで確認することができる。

```
>library(relimp)
>showData(x)  # データフレーム x を別ウィンドウで表示
```

## 3.1　例題 1-1: アヤメデータセットの探検

**レポート課題 1-1**

```
>library(MASS) :: data() の中に iris がないとき実行する
>iris
```

上記コマンドを実行して表示されるデータの、行ラベルと列ラベルを書き出し、その意味を示しなさい。データが多いので、適当に省略して説明すること。

```
>iris[1:4]
```

を実行すると、iris データフレームの 1 番目から 4 番目までの列の属性値が、行番号とともに表示される。また、


(c) Iris Virginica

図 1　アヤメデータセットの 3 種のアヤメ。外側の大きな花弁状のものが萼 (sepal、外花被ともいう)、萼に重なっている花弁状のものが花柱 (裏におしべがある)、萼の間にあるのが花弁 (petal、内花被ともいう)。

```
>iris$Species
```
を実行すると、Species の属性値 (クラス名) を要素とするベクトルが得られる。$記号は iris というデータフレームであることを指定する識別子である。
このコマンドを実行して出力される最後の行の　Levels: の後の文字列が、この列に含まれるクラス名のリスト要素を表示している。

```
>attach(iris)
```
を実行すると、“iris$” を省略できる。

```
>detach(iris)
```
で元に戻る。

すべての変数 (オブジェクト) は、何らかのタイプ (型) に属している。どのタイプに属しているかは、例えば、

```
>class(iris)
[1] "data.frame"
```
のように調べることができる。

**レポート課題 1-2**　iris データの各列のタイプとその意味を調べなさい。

データ処理や処理結果の表示を行う場合、クラスラベルのように長い文字列よりも、1 桁の数値や 1 文字でクラスを表した方がよい場合がある。そのような目的で、クラスラベルを剥がしてデフォルトの 1 文字や指定した 1 文字に置き換える関数が用意されている。

```
>unclass(iris$Species)
```
を実行すると、クラスラベルが整数値コードに置き換えられて表示される。データフレーム作成時に整数値コードにクラスラベルがアサインされている。

**レポート課題 1-3**

```
>c("a","b","c")[unclass(iris$Species)]
```
を実行し、このコマンドの動作を説明しなさい。

散布図とは、2 つの属性値をそれぞれ横軸と縦軸の座標値と見なして、2 次元上に各データ点を配置したものである。例えば、以下のコマンドを入力すれば、

```
>x<-runif(10)
```

```
>y<-runif(10)
>plot(x,y)
```
$(x, y)$ を座標として 10 個のデータが配置された図 2 のような散布図が得られる。

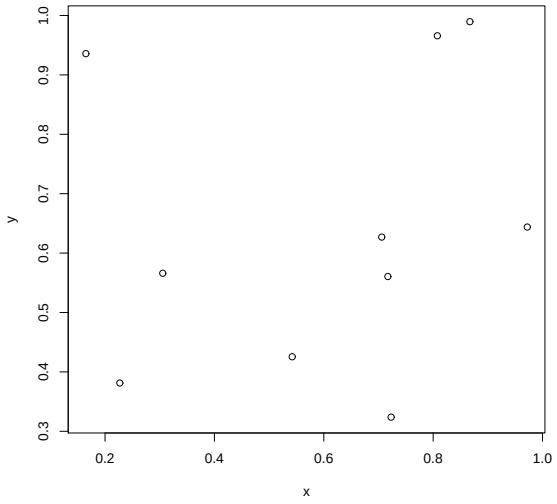



図 2　2 つの属性を持った 10 個のデータの散布図の例

LaTeX などでレポートを作成する場合、eps ファイルがよく使用される。図が表示された状態で

```
>dev.copy2eps(file="filename.eps")
```

というコマンドを入力すれば、filenmae.eps というファイルが作られる。

また、表示された図を右クリックすると”ポストスクリプトに保存”タブが出るので、それをクリックしても eps ファイル形式で保存することができる。他のファイル形式を選択することもできるので、word を使ってレポートを書く場合便利である。

アイリスデータのように属性が 2 つ以上ある場合は、

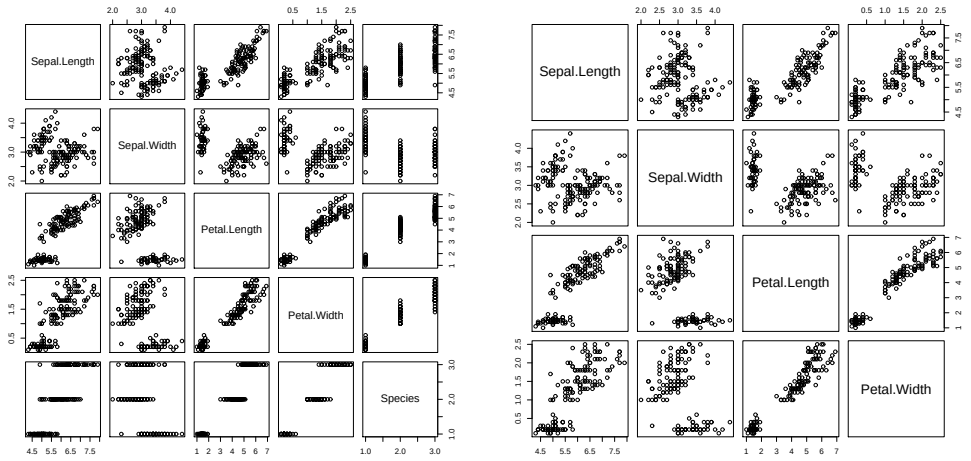
```
>pairs(iris)
```

を実行すれば、図 3(a) に示したように、全ての属性を対にした散布図の配列が得られる。

萼と花弁の数値属性の対のみの散布図を表示するためには、

```
>pairs(iris[1:4])
```

のようにすればよい。図 3(b) に結果を示した。




(a) (b)

図 3　アイリスデータの全ての属性間の散布図

ただし、図 3 では、アヤメの種類によるデータの区別ができない。アヤメの種類を色の違いで表示するためには、

```
>pairs(iris[1:4], pch=21, bg=c("red", "green3",
"blue")[unclass(iris$Species)])
```

のようにすればよい。

**レポート課題 1-4**

上記コマンドを実行した結果を提出するとともに、上記コマンドで用いたオプション、pch=21 と bg の意味を調べなさい。

**レポート課題 1-5**

上で得られた散布図から、アヤメの種類を識別するときに最も誤りが少なくなりそうな特徴を選び、その理由を説明しなさい。

**レポート課題 1-6**

data() コマンドで iris 以外のデータを一つ探し、レポート課題 R1-5 のような散布図を作成し、利用したデータや属性について説明しなさい。