

## Rでのデータの視覚化(1)

札幌学院大学社会情報学部教授

金明哲 (Jin Mingzhe)

■中国生まれ。総合研究大学院大学数物研究科統計科学専攻博士後期課程修了。博士(学術)。1995年より札幌学院大学。



われわれ人間は五感を通じて情報を得る。その中の大半は視覚によるものである。データをグラフで表現することは、データの中に潜んでいる情報を視覚的に考察する手助けとなる。Excelのような表計算ソフトには多様なグラフ作成機能がある。Rで計算加工したデータをテキストファイルとして出力し、Excelの機能を用いてグラフを作成することもできるが、直接Rでグラフを作成することが便利な場合もある。またExcelで作成できないグラフをRで簡単に作成することもできる。

### 1. 棒グラフ

グラフを作成するためにはデータが必要である。Rの中には、多くのデータセットが用意されている。Rに用意されたデータを用いるためには、保存されたデータをRに呼び出す必要がある。データを呼び出す関数は**data**である。

VADeathsというデータは、年齢50~74を5段階に分け、その死亡率を田舎(rural)、都会

(urban)のそれぞれ男(male)女(female)別に分けた5行4列の行列型データである。このデータをRに呼び出すには次のコマンドを実行する。さらに、コンソールにVADeathsと入力し、[Enter]キーを押すとデータが返される。

```
>data(VADeaths)
>VADeaths
```

	Rural	Male	Rural	Female	Urban	Male	Urban	Female
50-54		11.7		8.7		15.4		8.4
55-59		18.1		11.7		24.3		13.6
60-64		26.9		20.3		37.0		19.3
65-69		41.0		30.9		54.6		35.1
70-74		66.0		54.3		71.1		50.0

このデータを用いて、棒グラフを作成してみる。棒グラフの作成は、関数**barplot**を用いる。次のようにコマンドを入力し実行すると、次ページ図1のような棒グラフが作成される。

```
>barplot(VADeaths)
```

図1は1列のデータを1本の棒に表示している。ここで、各列のデータの順序と棒グラフにおけるデータ順序は逆になっていること

に注意されたい。つまり第1行のデータを棒の最も下の方に、第5行のデータを最も上の方に配置している。

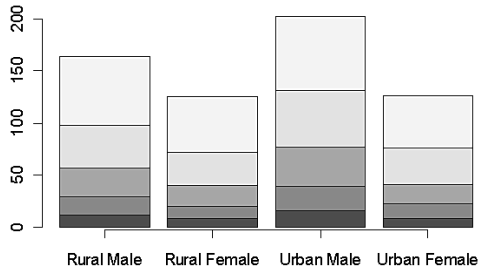


図1 棒グラフの例

コマンド

```
>barplot(VADeaths,beside=TRUE)
```

を実行すると、図2のような棒グラフが作成される。このbesideは棒グラフを横に並べるかどうかを指定する引数である。引数beside=TRUEにすると、各列のデータを行ごとに棒を横に並べて作成する。引数besideを用いないか、あるいはbeside=FALSEにすると図1のような図が作成される。

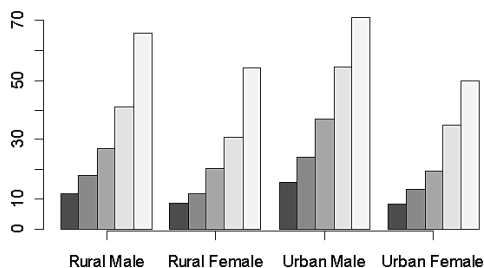


図2 棒グラフの例

棒の色は自由に調整することが可能である。例えば、コマンド

```
>barplot(VADeaths, beside = TRUE, col = c("lightblue", "mistyrose", "lightcyan", "lavender",
```

```
"cornsilk"), legend = rownames(VADeaths))
>title(main = "Death Rates in Virginia", font=5)
```

を実行すると図3のような棒グラフが作成される。ここでは4つの引数、データ、beside、col、legendが用いられている。

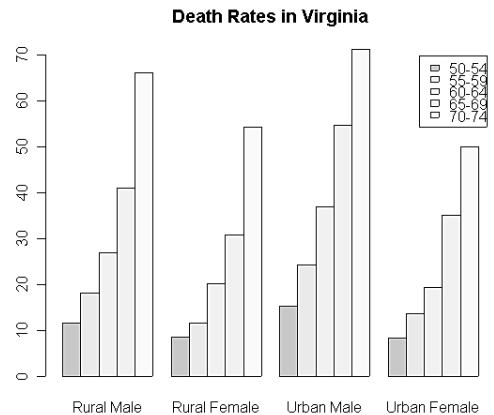


図3 棒グラフの例

引数colは棒の色を設定する引数である。ここでは列の数に相当する5色を指定している。色を文字列で指定する場合は、色の名前を“”で囲む必要がある。また、ここでは“lightblue”、“mistyrose”、“lightcyan”、“lavender”、“cornsilk”という色を指定したが、自分で好きな色を自由に指定することができる。ただし、色はRで指定されている色でなければならない。Rで指定されている色の種類はコマンド

```
>colors()
```

を実行することで見る事ができる。表示された657色の中から好きな色を指定すればよい。

引数legendは各棒に用いた色とデータを対応付けた凡例を作るための引数である。このrownamesは行の名前を用いることを意味する。

関数titleはグラフのタイトルを書く関数で、

表1 barplotの書式と主な引数

書式	barplot(x, beside=, horiz=, col=, legend=)
x	用いるデータ
beside=	TRUEの場合、1列を1組とし、各データを一つの棒に描くFALSEの場合、1列のデータを一つの棒に描く
horiz=	TRUEの場合には横棒、FALSEの場合は縦棒
col=	棒の色 (red/赤,blue/青,yellow/黄,green/緑,purple/紫,cyan/青緑,pink/ピンク,lightblue/薄青,lightcyan/薄青緑,lightgreen/薄緑,mistyrose/薄ピンク,lavender/ラベンダー,cornsilk/トウモロコシ毛の色,……色は文字列ではなく数値で指定することもできる。0は白、1は黒、2は赤、3は緑、4は青、5は青緑、6はピンク、7は黄……)
legend=	棒の説明、棒の塗りわけ方、範囲を指定する

メイン (main)、サブ (sub)、横軸 (xlab)、縦軸 (ylab) のタイトルの作成や、その文字の色 (col)、フォント (font)、サイズ (cex) などを指定することが可能である。

## 2. 円グラフ

ここではすでに入力した果物の売上の割合データ sales2 を用いることにする。

```
> sales2
  Cherry Apple Grape Banana Other
A    15    20    25     10    30
B    10    25    20     25    20
```

果物の売上を円グラフにしたとき、各果物の色をCherryはvioletred1、Appleはgreen3、Grapeはpurple、Bananaはyellow、Otherはcyanで表すことにする。そのためには、次のように各果物の色に対応する文字列ベクトルを作成しておくことと便利である。

```
> sales.col<-c("violetred1",
               "green3", "purple", "yellow", "cyan")
```

円グラフ (パイグラフ) を作成する関数は **pie** である。次のコマンドにより図4の円グラフが作成される。関数の中の引数radiusは円の半径に関する引数で、数値が大きければ、円が大きくなる。

```
> pie(sales2[1,],col=sales.col, radius=1)
```

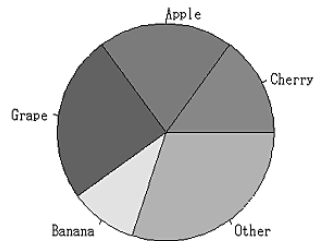


図4 円グラフ

次のような引数

```
col=gray(seq(0.1,1,length=データの数))
```

を関数pieに用いると円グラフを白黒で作成することができる。また、引数「density = 値」を用いると円グラフの各部分に斜線を引いた図を作成することができる。「値」が大きいほど斜線の間隔が狭い。斜線の角度は引数「angle = 値」を用いて調整する。

## 3. ヒストグラム

データを小さいものから順に並べ、階級ごとにまとめ、棒グラフで表現したものをヒストグラム (histogram) という。ヒストグラムはデータの全体状況をつかむのに有効である。ヒストグラムに適したデータは、長さ、重さ、

速度など連続な量的データである。

ここでは、Rに用意されているiris（アヤメ花、花菖蒲）というデータを用いてヒストグラムの作成について説明する。irisのデータは、setosa、versicolor、virginicaという3種類の品種のアヤメの花について、各種類それぞれ50標本の、がく片の長さ（Sepal.Length）、がく片の幅（Sepal.Width）、花弁の長さ（Petal.Length）、花弁の幅（Petal.Width）を計測した150標本のデータである。Rに用意されているirisのデータの形式は、150行5列のデータフレーム型である。第1列から第4列までが、それぞれがく片の長さ、がく片の幅、花弁の長さ、花弁の幅、第5列は品種のラベルである。1行から50行までがsetosaという品種で、51行から100行までがversicolorという品種で、101行から150行までがvirginicaという品種のデータである。データirisの形式を表2に示す。

表2 irisデータ形式

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
...	...	...	...	...	...
51	7	3.2	4.7	1.4	versicolor
...	...	...	...	...	...
101	6.9	3.3	6	2.5	virginica
...	...	...	...	...	...
150	5.9	3	5.1	1.8	virginica

コマンドdata(iris)を実行することにより、パッケージの中のirisデータが呼び出される。Rのコンソールにiris[1,]と入力し、[Enter]キーを押すと、第1行のデータが返される。同じくコマンドiris[51,]、iris[101,]を実行すると表2の51行目、101行目に相当するデータが返されることが確認できる。データが準備できたので、ヒストグラムの作成に戻る。

R上でヒストグラムを描く関数はhistである。データirisの中の品種setosaの第2変数Sepal.Width（がく片の幅）のヒストグラムを作成してみる。品種setosaは、irisデータの1行から50行までである。品種setosaの第2変数のヒストグラムは、コマンド

```
> hist(iris[1:50,2])
```

で作成される。その結果を図5に示す。ヒストグラム図5から、品種setosaのがく片の幅は3～3.5の周辺に集中していることがわかる。図3の5つの棒の幅となる範囲、2～2.5、2.5～3、3～3.5、3.5～4、4～5を階級という。

このように作成したヒストグラムでは、データの値が境界線の値と同じ場合は、右の棒（階級）にカウントされる。例えば、図5では値が2.5の場合は、階級2.5～3に属する。

Histogram of iris[1:50, 2]

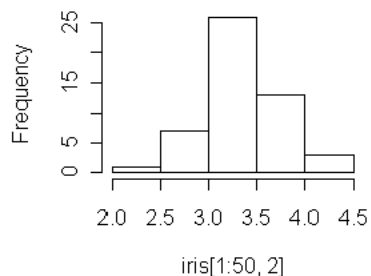


図5 setosaの第2変数のヒストグラム

ヒストグラムの各棒に頻度を表したり、棒に色付けをしたりすることも可能である。例えば、コマンド

```
> hist(iris[1:50,2], col="gray", labels = TRUE)
```

により、次ページ図6に示すヒストグラムが作成される。colはヒストグラムの色、labels

は度数を指定する引数である。

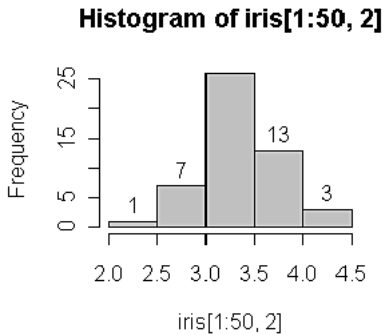


図6 色と度数を加えたヒストグラム

ヒストグラムでは、引数breaksを用いて各自が階級を自由に設定することができる。引数breaksの階級設定は、各棒の境値のベクトルを用いる。ただし、階級の最小値と最大値はデータの範囲外でなければならない。

例えば、がく片の幅のヒストグラム階級を1~2、2~3、3~4、4~5にしたいときには、次のようにbreaks=c(1,2,3,4,5)を引数(あるいはbreaks=c(1:5))とする。次の書式により作成したヒストグラムを図7に示す。

```
> hist(iris[1:50,2],breaks=c(1:5),col="blue")
```



図7 階級の幅を1としたヒストグラム

作成したヒストグラムに関連する情報を文字列として出力したいときには引数

plot=FALSEを用いる。度数はcounts、階級の区分点はbreaksという名でその値を返す。

#### 4. 箱ひげ図 (box-whisker plot)

データ解析では、データの中心、散らばりの具合、異常なデータなどを考察するために、図8のような長方形(箱)に直線(ひげ)をつなげたグラフをよく用いる。このグラフを箱ひげ図という。

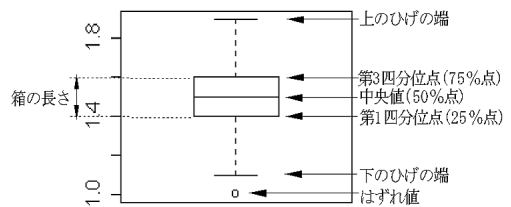


図8 箱ひげ図の構造

箱ひげ図では四分位数を用いる。四分位数とはデータを小さい値から大きい順に並べ、その範囲を4等分した場合、小さい値から第1等分と第2等分の境である25%点を第1四分位数、第2等分と第3等分の境である50%点を第2四分位数、第3等分と第4等分の境である75%点を第3四分位数という。

箱ひげ図の箱の長さは、第3四分位数から第1四分位数の値を引いた値である。ひげの長さは、一般的に箱の長さの1.5倍とする。この1.5倍は絶対的なものではない。Rでは自由に設定することができるが、デフォルト(初期に設定されている値)のひげの長さは箱の長さの1.5倍である。ひげの上(下)端は、箱の長さの1.5倍以内にある最大(小)値である。ひげの端の外側にある値を外れ値、あるいはアウトサイド値とする。

箱ひげ図を用いてirisデータのがく片の長さ

(変数1) について、3種類のアヤメを比較してみよう。コマンド

```
>boxplot(iris[1:50,1],iris[51:100,1],iris[101:150,1],col="gray",names=c("S", "C", "V"))
```

を実行すると、図9に示す箱ひげ図が作成される。引数colは箱の色、namesは各箱ひげ図のラベルを指定する引数である。

図9からそれぞれの、がく片の長さの特徴が概観できる。長さが最も短いのがsetosaという品種で、その次がversicolorで、最もがく片が長いのはvirginicaという品種である。かつvirginicaの中には、一つのデータが特に小さいことも読み取れる。

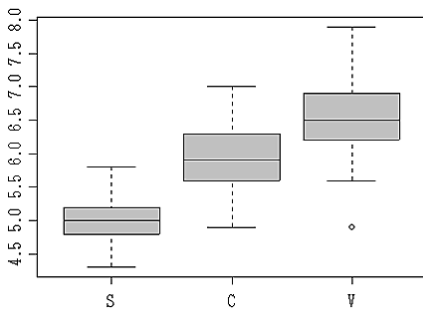


図9 irisのがく片の長さの箱ひげ図1

## 5. 日本語の環境設定とグラフでの表示

Rで日本語を扱うためには特別な手続きが必要である。Rの新しいバージョンR1.8.0がリリースされたので、WindowsのR1.8.0を例とし、日本語環境設定の手順を説明する。

まず、R1.8.0をダウンロードし、インストールする。次に以下の手順で日本語環境を設定する。

- ① <http://r.nakama.ne.jp/R-1.8.0/binary/win32> からR-1.8.0-L10Npack\_YYYYMMDD.tar.gzをダウンロードし、解凍する。ファイルの名

前のYYYY、MM、DDはファイルを公開した年月日である。

- ② 解凍されたフォルダの中のbin、etc、libraryの三つのフォルダをコピーし、Rがインストールされているフォルダ(例えば、C:\Program Files\R\Rw1080)の中のbin、etc、libraryに上書きする。

上記の作業で、日本語が使えることになる。

R1.8.0から今まで使用しているR1.7.1の作業画面を読み込むことができる。

- ① R1.8.0を起動し、Rのメニューの「File」⇒「Load Workspace」をクリックすると「Select image to load」ダイアログボックスが開かれる。

- ② 「Select image to load」ダイアログボックスの「ファイル場所 (I)」にrw1071がインストールされている場所を指定し、「ファイル名前 (N)」にRDataファイルを指定し、「開く (O)」ボタンを押す。

円グラフを作成する際に用いたsales2の果物名を日本語にした棒グラフ作成を例とし、日本語の使用例を次に示す。日本語を用いる際には、日本語の表記以外の「”」や「,」などの記号は半角文字にしなければならない。

```
> ラベル<-c("さくらんぼ", "林檎", "葡萄", "バナナ", "その他")
> colnames(sales2)<-ラベル
> barplot(t(sales2),horiz = T,col = sales.col,legend = colnames(sales2), xlim = c(0, 140))
> title(main = "果物売り上げの横棒グラフ")
```

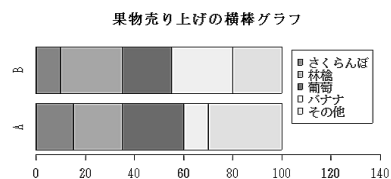


図10 irisのがく片の長さの箱ひげ図2