

Rでのデータの視覚化(3)

8. まざまな散布図

8.1 3変数の散布図

通常の2次元平面上の散布図は2変数を表現しているが、横軸と縦軸を2変数に対応させ、図形の形や大きさなどで1変数を対応させることで、2次元平面上で3変数を表すことができる。

Rでは関数 `symbols` を用いて3変数の散布図を作成することができる。Rには31本の切り倒された桜の木について3変数で測定した `trees` というデータがある。データ `trees` の3変数はそれぞれ、`Girth`(木の直径、単位はインチ)、`Height`(木の高さ、単位はフィート)、`Volume`(木の体積、単位はフィート)である。

関数 `symbols` を用いたデータ `trees` の `Height` を横軸、`Volume` を縦軸、`Girth` を散布図の点の直径とした3変数の散布図の作成過程と結果を示す。散布図から、木の直径、高さ、立方体の関係が読み取られる。

```
>data(trees)
>attach(trees)
> symbols(Height, Volume, circles=Girth/10, inches=FALSE, bg = c(1:31))
```

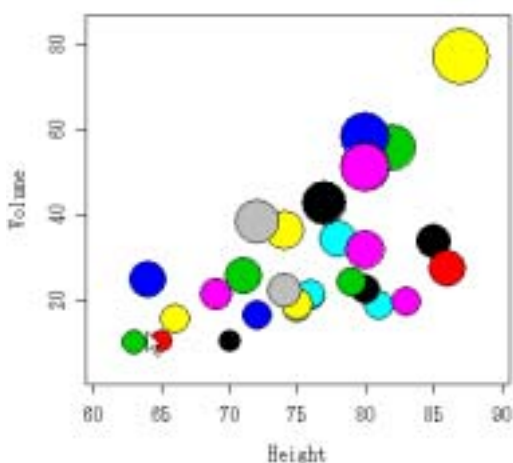


図 29 3変数の散布図

8.2 条件付散布図

条件付(conditioning)散布図は、ある条件が与えられたもとでの 2 変数の散布図である。R での条件付散布図を作成する関数は `coplot` である。関数 `coplot` の書式を次に示す。

```
coplot(縦軸の変数 ~ 横軸の変数 | 横軸の条件*縦軸の条件, data = データ)
```

例として、R の中の `iris` データの品種を前提条件、花弁の長さ(`Petal.Length`)を横軸、幅(`Petal.Width`) を縦軸にした散布図の作成コマンドとその結果を次に示す。

```
>data(iris)
>coplot(Petal.Width ~ Petal.Length | Species, data = iris)
```

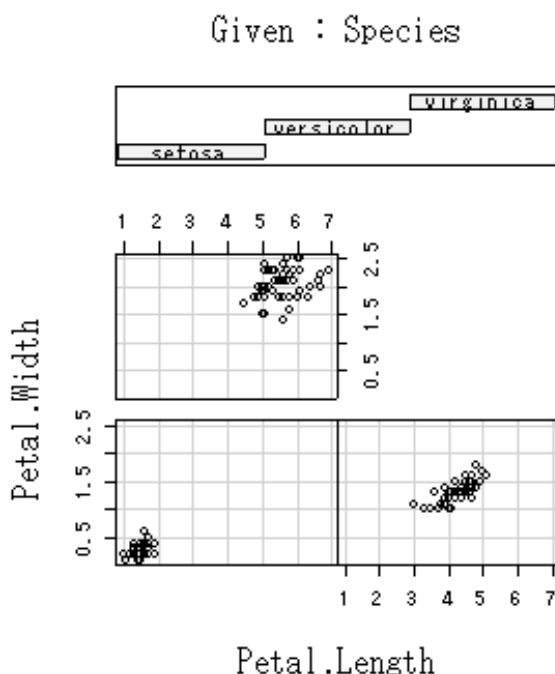


図 30 iris の条件付散布図 1

図 30 の下段の左辺は `setosa`、下段の右辺は `versicolor`、上段は `virginica` の散布図である。図から `setosa` のほとんどは $Petal.Width < 0.5$ 、 $Petal.Length < 2$ で、`versicolor` のほとんどは $1 < Petal.Width < 2$ 、 $3 < Petal.Length < 5$ で、`virginica` のほとんどは $1.5 < Petal.Width < 2.5$ 、 $5 < Petal.Length < 7$ であることがわかる。

次のコマンドのように条件(`Species`)を次のコマンドのように設定することにより、図 31 のような条件付散布図を作成することはできる。

```
>coplot(Petal.Width ~ Petal.Length | Species*Species, data = iris)
```

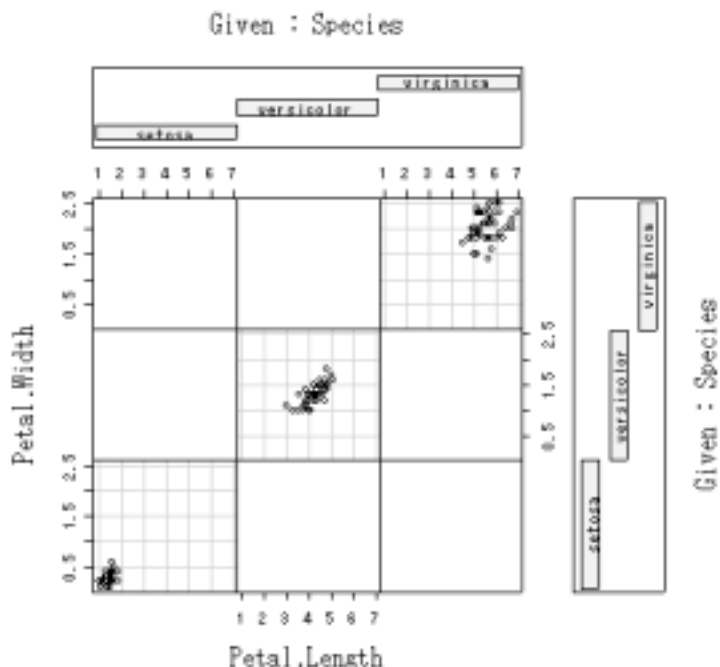


図 31 iris の条件付散布図 2

R に Prestige というデータセットがある。Prestige は 1971 年にアメリカで 102 業種について 6 項目分けて行った調査データである。データは 102 行 6 列のデータフレーム型である。紙面の都合により、次に用いる 3 つの変数のみ説明する。

education: 教育を受けた平均の年数

income: 1971 年の平均収入、アメリカドル

prestige: 職業に対するスコア

データ Prestige はパッケージ car の中に含まれている。もし、パッケージ car がインストールされていない場合は、まずインストールを行った上で、次のように読み込む必要がある。

```
>library(car)
>data(Prestige)
```

次のコマンドのように、関数 coplot に引数 panel = panel.smooth を加えることにより

データの近似曲線を追加することもできる。

```
>coplot(prestige~income|education, panel = panel.smooth, data=Prestige)
```

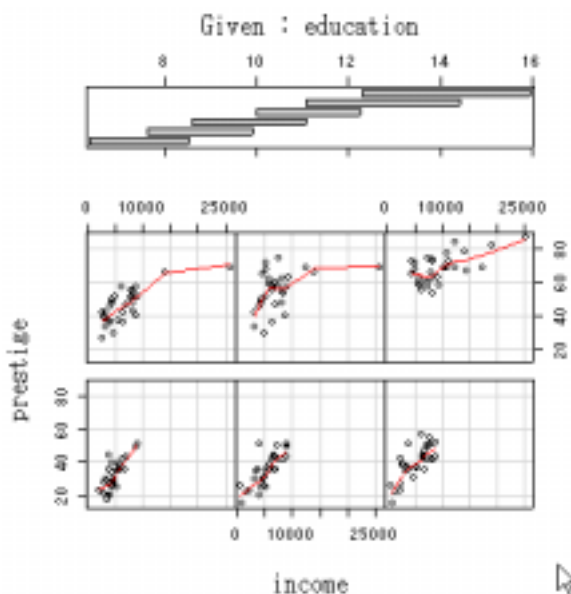


図 32 Prestige データの条件付散布図

この条件付散布図の条件は「education」である。作成された散布図は教育水準を自動的に6つのカテゴリに分け、それに対応する散布図を6つ返している。教育を受けた年数と散布図との対応関係は、散布図の左辺から右辺に、下段から上段の順に教育を受けた年数のバーの左から右へのカテゴリに対応されている。この条件付き散布図から、教育を受ける年数が長いほど職業のスコアが高く、かつ収入も高いことが読み取られる。

9. モザイクグラフ

モザイクグラフは、クロス表のようなセルの度数データを長方形で示す。その長方形の面積で、セルの値を示す。

Rでのモザイクグラフを作成する関数は**mosaicplot**である。もし、作成されたデータがテーブル型の場合は関数**plot** でモザイクグラフを作成することができる。

1912年4月10日、2千人以上の乗客を乗せ、イギリスのサウサンプトン港を出港し、アメリカのニューヨークに向かったタイタニック号が悲慘な海難事故を引き起こし、4月14日沈没したことは世界で知られている。その際の生還者と死亡者を乗員の等級、性別などに分けて整理したデータTitanicがRの中にある。

```
> data(Titanic)
```

```
> class(Titanic)
```

```
[1] "table"
```

```
> dim(Titanic)
```

```
[1] 4 2 2 2
```

上記のコマンドの実行結果からわかるように、データTitanicはテーブル型で、4層になっている。このテーブルのデータ構造は、配列の構造と同じであると考えても差し支えない。

表1 各層内の行の番号と内容の対応

1層	2層	3層	4層
1等 = 1	男 = 1	子供 = 1	死亡 = 1
2等 = 2	女 = 2	大人 = 2	生還 = 2
3等 = 3			
4等 = 乗務員			

データTitanicを用いて、モザイクグラフの作成についてに説明を行う。男の大人の死亡・生還データは、次のコマンドで呼び出すことができる。

```
> Titanic[,1,2,]
```

```
      Survived  
Class  No Yes  
1st   118  57  
2nd   154  14  
3rd   387  75  
Crew  670 192
```

またこのデータのモザイクグラフは次のコマンドで作成することができる。データTitanicはテーブル型であるので、関数plotを用いても同様なモザイクグラフを作成することができる。

```
> mosaicplot(Titanic[,1,2,],col=T)
```

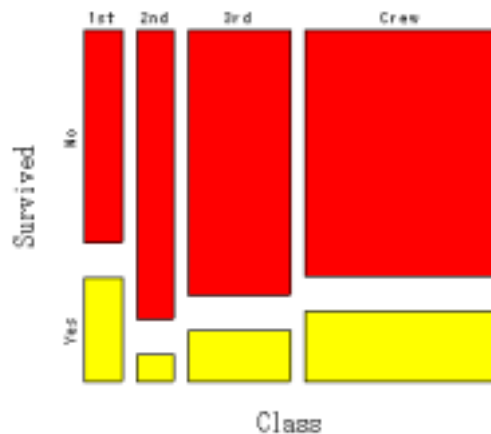


図 33 大人の男性乗員のモザイクグラフ

モザイクグラフは、長方形の面積を用いて各セルの値の大小を示す。例えば、図 33 では、縦軸の上の部分が生還できなかった割合、下の部分が生還できた割合で、縦の辺の長さを用いて乗客の等級内での生還・死亡の割合を示す。横軸は乗員の等級に関する情報で、長方形の横辺の長さを用いて、乗員の各等級の割合を示す。乗員を等級別に見ると、2 等級の生還率が最も低いことがわかる。

タイタニックの全てのデータを用いたモザイクグラフを次に示す。

```
>mosaicplot(Titanic,col=T)
```

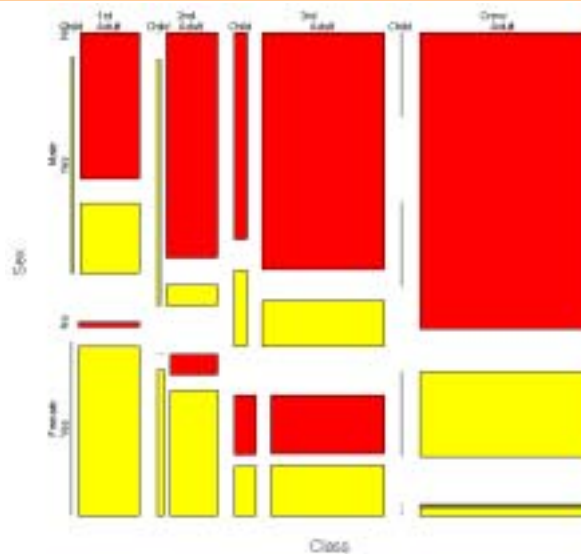


図 34 Titanic データのモザイクグラフ

上記のモザイクグラフをざっと見るだけで、女性の生存者の割合が男性より高いことがわかる。

10. 星図

星図は名前のとおりデータを星の形にしたグラフである。レーダーチャートのようなものである。星図は1つの星(多角形)がデータセットの1行(個体)を示し、中心から伸びている線が1つの変数を表す。ただし、Rでは各変数(列)は、最も大きいのを1になるように変換し、星図を作成する。星図では、作成された星の形から、データの特徴を視覚的に分析する。

ここでは longley という 1947 年から 1962 年までの 16 年間の 7 列の経済データを用いる。データの各列は物価調整済みで、1954 年を 100 としている。第 1 列から GNP、GNP、失業者数、軍隊人員数、14 歳以上の人口数、年度、雇用数の順である。関数 stars を用いた星図の作成とその結果を示す。関数 stars に用いた引数 key.loc = c(0,6) は凡例の座標の設定である。

```
> data(longley)
> stars(longley, key.loc = c(0, 6))
```

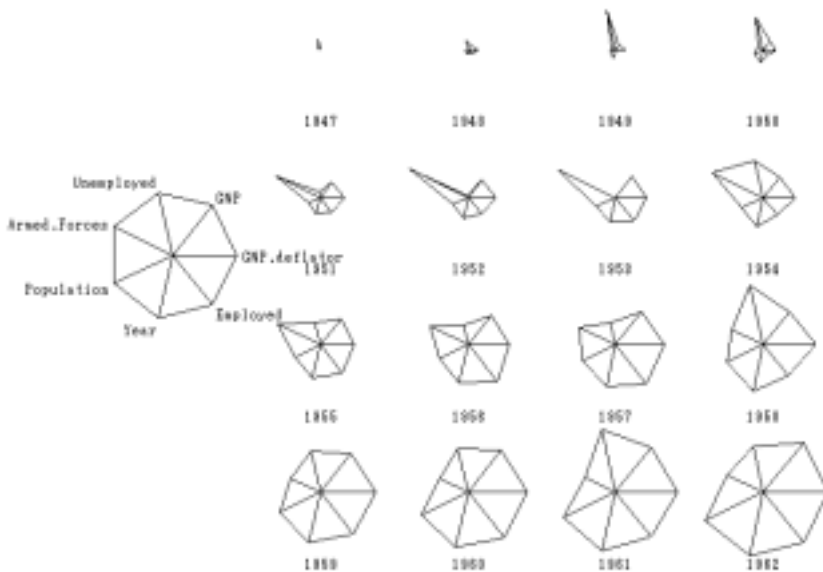


図 35 longley データを用いた星図の例 1

関数 stars に引数 full = FALSE を次のように用いると星を上半分になるように描く。

```
> stars(longley, key.loc = c(0, 6), full = FALSE)
```

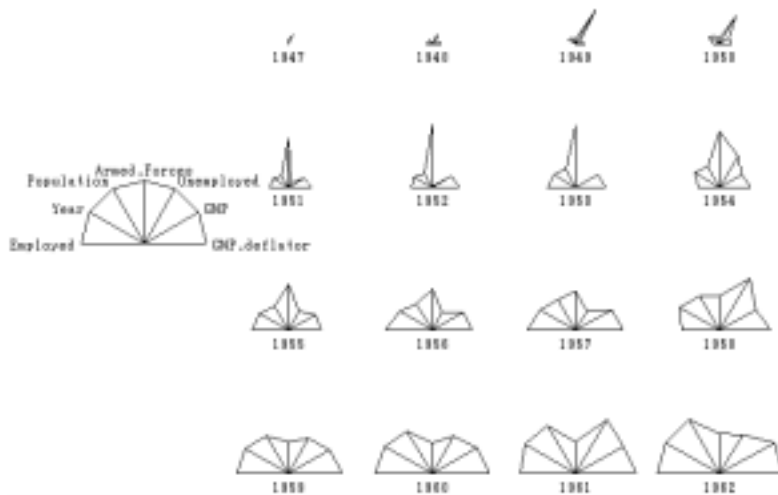


図 36 longley データを用いた星図の例 2

星図の関数に引数 `draw.segments = T` を加えることにより、着色した図 30 のような図を作成することができる。色は引数 `col.segments` に色を指定するベクトルを付値し、好みの色を用いることができる。

```
>stars(longley,key.loc = c(0, 6),draw.segments = T)
```

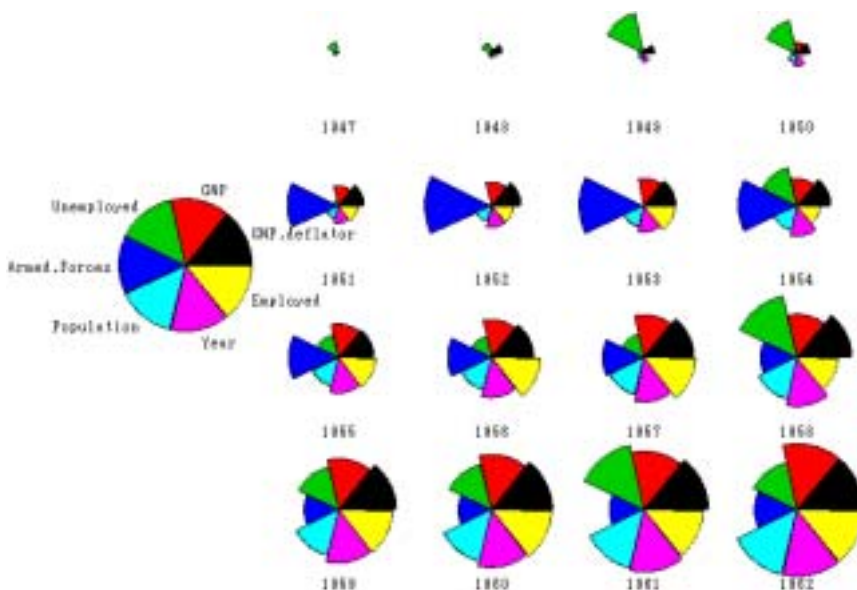


図 37 longley データを用いた星図の例 3

11. チャノフの顔形グラフ

チャノフの顔形グラフはチャノフ(H.Chernoff)が 1973 年に発表した多変数データの変数を人間の顔の各部位に対応付けたグラフである。チャノフの顔形グラフの狙いは、顔の表情からデータの特徴を読み取ることである。

R の本体には、チャノフの顔形グラフを作成する関数が用意されていない。群馬大学社会情報学部の青木繁伸教授は、チャノフ顔形グラフ作成のプログラムを <http://aoki2.si.gunma-u.ac.jp/R/face.html> で公開している。

上記のサイトのプログラムをコピーし、R のコンソール上に貼り付けるとチャノフの顔形グラフ作成関数 `face.plot` が利用可能になる。関数 `face.plot` に用いるデータは、事前処理が必要である。その処理を行うプログラム [face.data](#) は同じのページにリンクが張られている。`face.plot` の場合と同じくコピーし R のコンソールに貼り付けてください。

関数 `face.plot` では最大 18 変数を表示することができる。変数と顔の部位との対応関係を表 5 に示す。

変数の数が 18 未満の場合は、ゼロを加え 18 変数になるようにし、関数 `face.data` を用いてチャノフの顔形グラフ作成関数 `face.plot` 用のデータセットを作成する。

R に弁護士によるアメリカ合衆国最高裁判官を評価したデータ `USJudgeRatings` がある。このデータは 43 個体(行)12 変数(列)である。

表 5 変数と顔の部位との対応関係

x_1	: 上半顔の大きさ
x_2	: 上半顔と下半顔の接続位置
x_3	: 顔の長さ
x_4	: 上半顔の楕円の離心率
x_5	: 下半顔の楕円の離心率
x_6	: 鼻の長さ
x_7	: 口の位置
x_8	: 口の曲率
x_9	: 口の幅
x_{10}	: 目の位置
x_{11}	: 目の中心の離れ度合
x_{12}	: 目の傾き
x_{13}	: 目の楕円の離心率
x_{14}	: 目の幅の半分
x_{15}	: 瞳の位置
x_{16}	: 目から眉の距離
x_{17}	: 眉の傾き
x_{18}	: 眉の長さ

全ての個体のチャノフの顔形グラフを示す紙面がないので、先頭の4つを用いて、顔形グラフを作成する手順を示す。

このデータは12変数であるので、各個体に6つのゼロを追加する。このゼロは、データベクトルのどの部分に入れるかに関しては自由である。説明の便利のため、最後の6つの変数 $x_{13} \sim x_{18}$ をゼロとする。よって作成された複数の顔形グラフの目の楕円の離心率、目の幅の半分、瞳の位置、目から眉の距離、眉の傾き、眉の長さは同じとなる。

```
>data(USJudgeRatings)
>da1<-USJudgeRatings[1:4,]
>da2<-matrix(0,4,6)
>da3<-cbind(da1,da2)
>da4<-face.data(da3)
>par(mfrow=c(2,2))
>for(i in 1:4)face.plot(da4[i,])
```

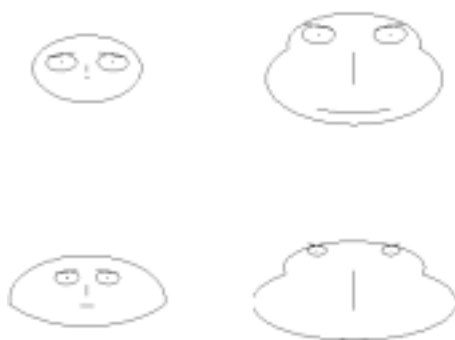


図31 チャノフの顔形グラフ

チャノフの顔形グラフはユニークであるが、変数と顔の部位の対応を換えると顔の形や表情が変わるので、変数と顔部位の対応の調整が大変である。