23日目：階層的クラスター分析

本日は階層的クラスター分析をやってみます。SPSSでやると，結構表示が遅い感じがするのですが，Rはサクサクと結果を出してくれます。

クラスター分析は，因子分析の方法と似ていて，様々な手法があります。距離の計算の仕方しかり，クラスタリングの方法しかり。いろいろと組み合わせることができますし，クラスター数をいくつにするかも特に基準があるわけではありません。そのため，やはり複数の結果を比較検討して決めるという手間がかかる分析です。

今回は，データとしてsam3.xlsxにある，「知識欲」「不可欠さ」「本好き」の3変数を使って，対象をクラスターに分けてみます。

Xという名前でデータを読み込んだら，以下のようにして変数を取り出し，標準化して，データフレーム形式にしておきます。

```
x0 <- c("知識欲","不可欠さ","本好き")
x1 <- x[x0]
x2 <- scale(x1)
x2<- data.frame(x2)
```

間隔尺度であれば，基本的には（単位に特に意味がないなら）標準化したものを使うことが推奨されるようです。

次に距離行列を求めます。

```
xd <- dist(x2, method="euclidean")
```

distは，距離行列を求める関数です。カッコの中は，ファイル名と，method=で求める方法を指定します。"euclidean"は，ユークリッド距離のことです。SPSSでは，「測定方法」で，「ユークリッド距離」と，「平方ユークリッド距離」を選択できますが，distでは「平方ユークリッド距離」を直接計算できないので，後で二乗してやります。

ヘルプを見ると，ユークリッドの他には"maximum"，"manhattan"，"canberra"，"binary"，"minkowski"の方法が選べるようです。間隔尺度のデータであるなら，基本的には，ユークリッドでよいのではないでしょうか…。

行列が準備できたらクラスター分析を実行させます。関数は，hclustです。カッコの中には，距離行列とクラスター分析の方法を指定します。なお，先に平方ユークリッド距離を

使う場合は，後で二乗するということを書きましたが，ここで二乗しています。方法は一般的なウォード法（method="ward"）を使っています。

```
clus1 <- hclust(xd^2, method="ward")
```

方法には"ward"の他に，"single"，"complete"，"average"，"mcquitty"，"median"，"centroid"，"complete"（これがデフォルト）が指定できるようです。

なお，これは余談かもしれませんが，以前のSPSSは，ウォード法（"ward"）や重心法（"centroid"），メディアン法（"median"）では，距離行列に平方ユークリッド距離のみが用いられていました（少なくともSPSSのver.4のマニュアルにはそのように記載があります）。またwebを見ても，（当然のように）そのように記載されている場合があります。ウォード法の場合，平方ユークリッド距離でなければならないのか否か…が気になりますが…わかりません…

結果の表示ですが，

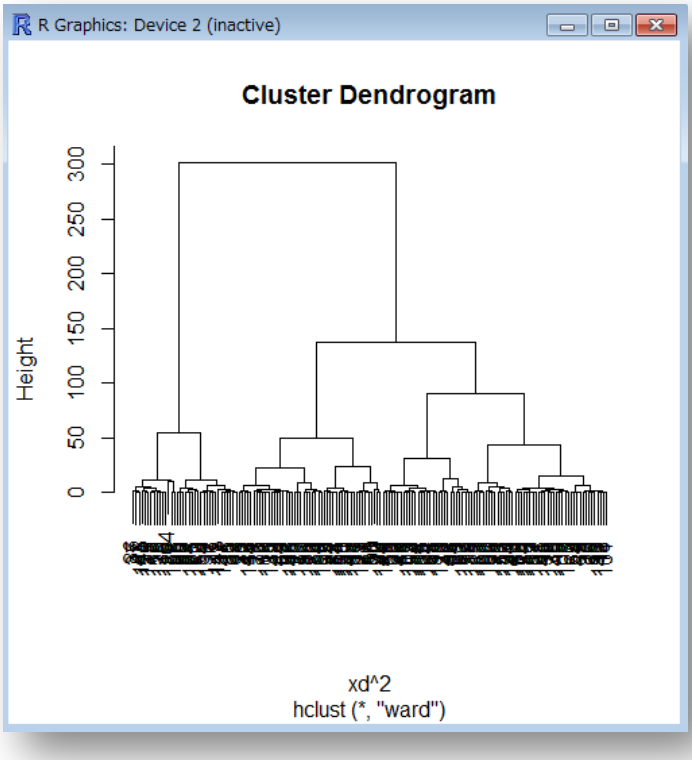
```
clus1
```

で中身を見ると，極めてそっけなく，クラスター分析の方法，距離行列，人数のみを返してきます。

デンドログラムを表示させるには…

```
plot(clus1)
```

これで右図のようなデンドログラムを作成してくれます。これを参考にしながら，いくつのクラスターを抽出するか悩みましょう。



しかし，悩むといっても，この図とにらめっこしていても，それぞれのクラスターの特徴はまったくわかりません。実際に取り出して，特徴を比較してみなければはじまらないでしょう。

そこで，クラスター数を指定して，それぞれのクラスターに特定の番号

を付し，クラスター間の異同を検討してみます。

```
x2$cl1 <- cutree(clus1, k=4)
```

cutree は，デンドログラムを指定する位置で切断し，各クラスターに番号を振ってくれます。カッコの中は，クラスター分析の結果と，k=で取り出すクラスター数を指定します。この例であれば，４つのクラスターを抽出し，分類番号を x2 に cl1 という変数名で保存しなさいという意味になります。

ついでに，３つを抽出する

```
x2$cl2 <- cutree(clus1, k=3)
```

もやっておいて結果を比較してみましょう。

これらを実行した後，cl1 と cl2 のクロス表を作成してみると右のようになりました。

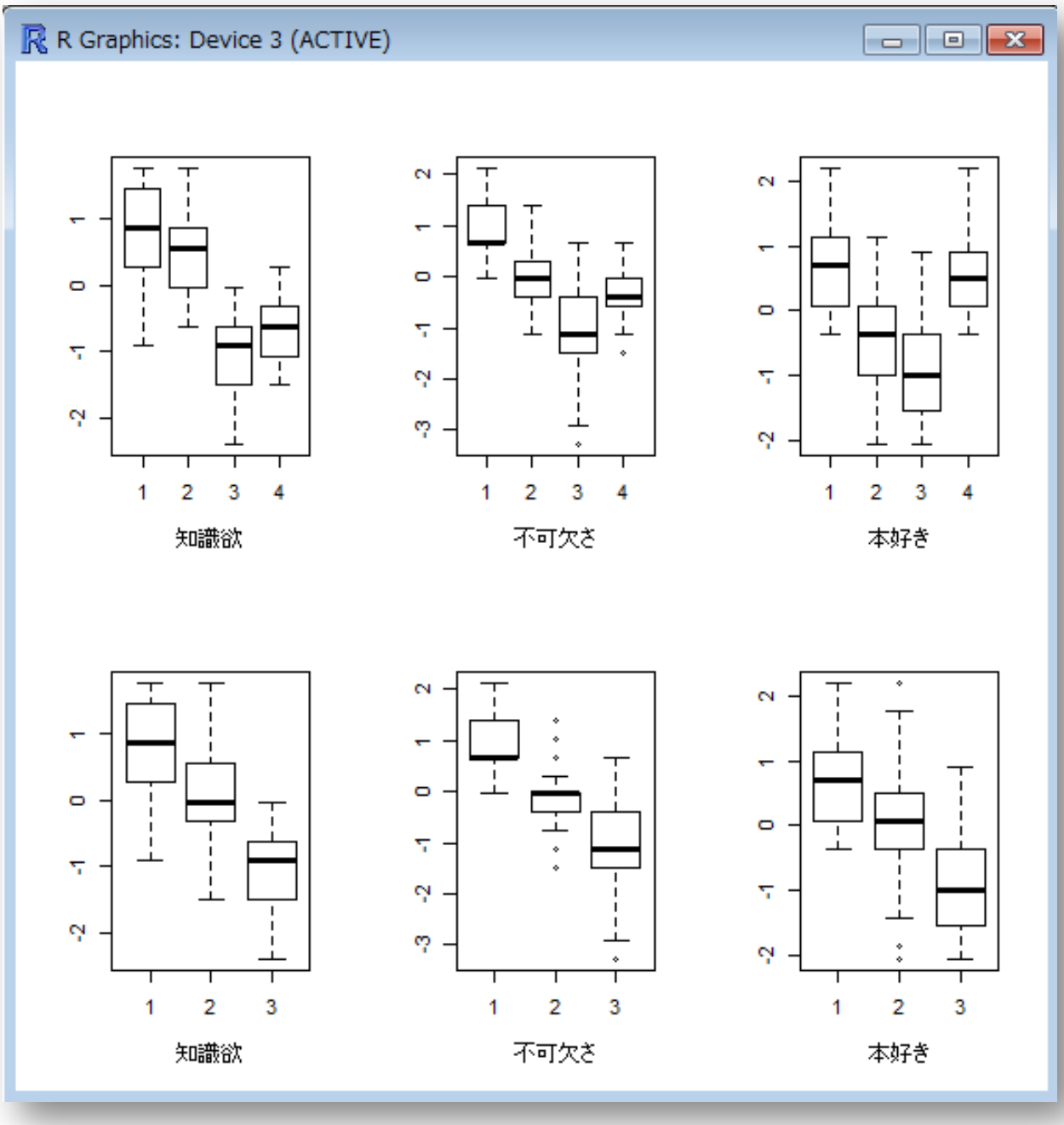
```
> table(x2$cl1,x2$cl2)

    1  2  3
1  49  0  0
2   0 46  0
3   0  0 36
4   0 31  0
```

３クラスターを抽出した場合の２が，４クラスターを抽出した時の２と４に分かれていることがわかります。クラスターの番号は，単にデンドログラムの右からとか，左から順にふられているわけではないようです。

さて，それぞれのクラスターの特徴把握ですが，もちろん describe.by などでクラスターごとの平均を計算し，エクセルにコピペしてグラフ化することができます。またR上でも，以前紹介した plotmeans や boxplot を使って概略を把握することができます。

boxplot を使って，３つのクラスターと４つのクラスターの場合を比較できるようやってみました。ここに貼り込むと図が小さくなりましたが，９日目に紹介した方法で，横を画面いっぱいまで広げて，３つのグラフを横に並べて表示してみました。なお，xlab=で各グラフの下に表示する内容を指定できます。

```
par(mfrow=c(2,3))
boxplot(x2$知識欲 ~ x2$cl1, xlab="知識欲")
boxplot(x2$不可欠さ ~ x2$cl1, xlab="不可欠さ")
boxplot(x2$本好き ~ x2$cl1, xlab="本好き")
boxplot(x2$知識欲 ~ x2$cl2, xlab="知識欲")
boxplot(x2$不可欠さ ~ x2$cl2, xlab="不可欠さ")
boxplot(x2$本好き ~ x2$cl2, xlab="本好き")
```



３つのクラスターの２が，４つのクラスターの２，４に分かれたわけなので，これだけを見ると，４つの方がよさそうです。

こんな作業を繰り返しながら，適切なクラスターを抽出していくことになります。