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。ひとつの変数を使って，対象者を上下位群に分けるとか，Yes／No 群に分けるといったこともありますが，クラスター分析は，複数の変数への回答傾向から，対象者を分類しようとする方法です。クラスターとは，類似した傾向をもつ対象者群というような感じのものになります。

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，以下のようにして利用する変数のみを取り出し，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)
x.c <- 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(x.c, 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。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，ウォード法ではユークリッド距離ではなく，「平方」ユークリッド距離，すなわちユークリッド距離の 2 乗値を使うのが一

般的なようです。dist でmethod= "euclidean"で求めたものはユークリッド距離なので，ここにちょっと注意が必要です。方法は 2 つあります。ユークリッド距離のデータをそのまま投入するか，先に 2 乗してから投入するかです。それによって，method=で"ward.D2"を指定するか，"ward.D"を使うかが決まります。ユークリッド距離のデータそのままなら"ward.D2"を，平方ユークリッド距離を与えるなら"ward.D"を指定します。

```
clus1 <- hclust(xd, method="ward.D2")
```

もしくは

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

方法にはウォード法の他に，"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。それぞれの特徴を紹介するには力不足なので，自分で調べてください…

結果の表示ですが，

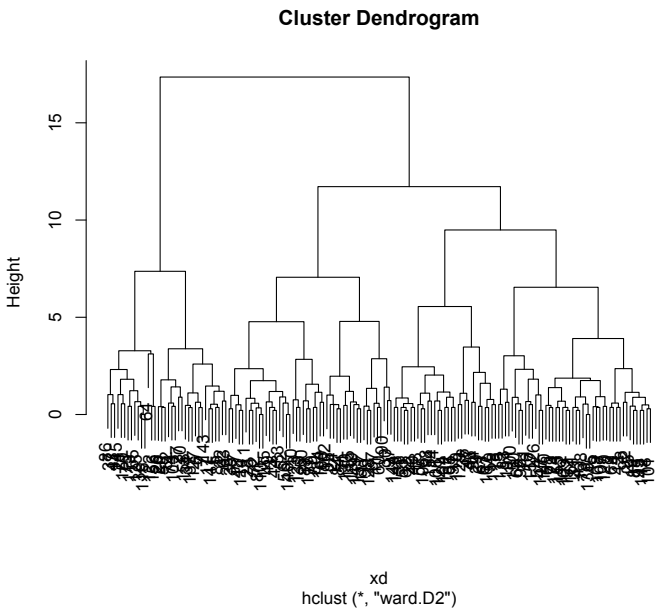
```
clus1
```

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（"ward.D2"を使った場合）。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。なお，先に"ward.D"と"ward.D2"のふたつがあることを指摘しましたが，デンドログラムを描くと，これらの結果は同じになりません。縦軸の数値が異なるようです。ただ，数値が異なるだけで，関係性は同じになります。そのため，以下の対象者の分類結果はどちらでも同じになります。



さて，ここからいくつのクラスターを抽出するか悩むといっても，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。

```
x$cl_4 <- cutree(clus1, k=4)
```

cutree は，デンドログラムを指定する位置で切断し（いくつのクラスターを取り出するか），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，その分類番号を，x に cl_4 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

```
x$cl_3 <- 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。

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。なお，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。

```
> table(x$cl_4, x$cl_3)

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

さて，それぞれのクラスターの特徴把握ですが，もちろん describeBy 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 を使って概略を把握することができます。

また箱ひげ図を書かせてみるのもよいでしょう。boxplot を使って，各クラスターの特徴や，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す。以下は，知識欲への回答の様子を x$cl_4 の４つのクラスター別に示せという指示になります。xlab= ， ylab=はｘ軸ｙ軸に付ける名前です。

```
par(family="Osaka")
boxplot(x$知識欲 ~  x$cl_4, xlab="クラスター番号", ylab="知識欲")
```

見方は，箱の上辺が第３四分位数，下辺が第１四分位数の位置を表します。そして箱の中のちょっと太い横線が中央値（平均値ではない！）です。そして箱から伸びた「ひげ」の先

は，ちょっとややこしい説明になります。たいていは，これを最大値と最小値と考えればよいのですが，この「ひげ」の先は，たとえば上側なら，第３四分位数から四分位範囲の 1.5 倍以内にあるデータのうちの最大値を示します。その外側に○がプロットされている場合はさらに注意が必要で，「ひげ」の端よりもさらに外れるようなデータがプロットされます。

少し図が小さくなりますが，３つのクラスターと４つのクラスターの３つのグラフを縦横に並べて表示してみます。

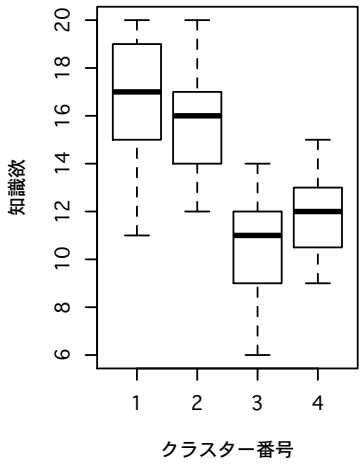
```
par(family="Osaka", mfrow=c(2,3))
boxplot(x$知識欲 ~ x$cl_4, xlab="クラスター番号", ylab="知識欲")
boxplot(x$不可欠さ ~ x$cl_4, xlab="クラスター番号", ylab="不可欠さ")
boxplot(x$本好き ~ x$cl_4, xlab="クラスター番号", ylab="本好き")
boxplot(x$知識欲 ~ x$cl_3, xlab="クラスター番号", ylab="知識欲")
boxplot(x$不可欠さ ~ x$cl_3, xlab="クラスター番号", ylab="不可欠さ")
boxplot(x$本好き ~ x$cl_3, xlab="クラスター番号", ylab="本好き")
```

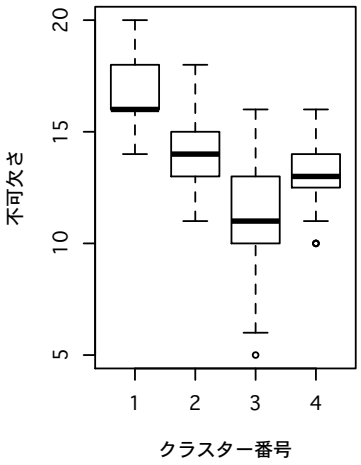
３つのクラスターと４つのクラスターの移動は，先に table で確認しました。３つのクラスターの２が，４つのクラスターの 2，4 に分かれているので，この情報も使いながら考えると，「不可欠さ」では，あまり違いませんが，その他では，クラスターを分けることで，それぞれの特徴がはっきりするようにも考えられます。

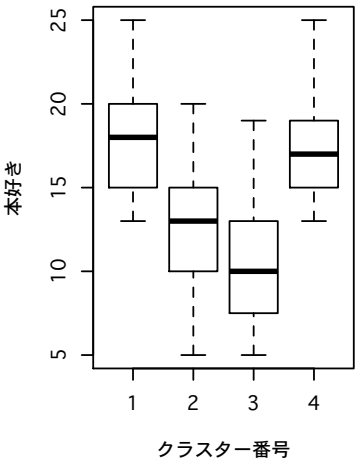
そして４つのクラスターを抽出した場合だと，1 に分類される対象者は「知識欲」「不可欠さ」「本好き」のいずれも得点が高い。対照的なのが 3 で，いずれも低い。2 と 4 は，「不可欠さ」は同程度であるものの，「知識欲」と「本好き」の高低で異なる。こういう特徴を持つ 4 群が抽出できたことになります。

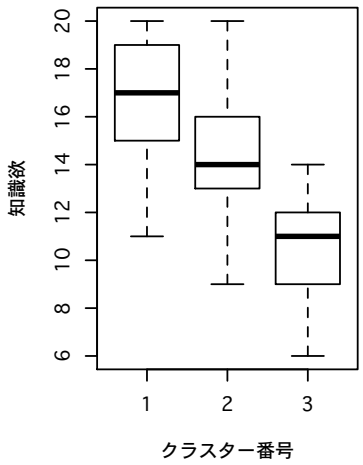
こんな作業を繰り返しながら，適切な数のクラスターを抽出していくことになります。

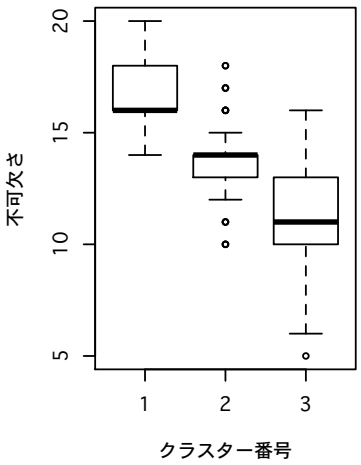
本日はここまでにします。

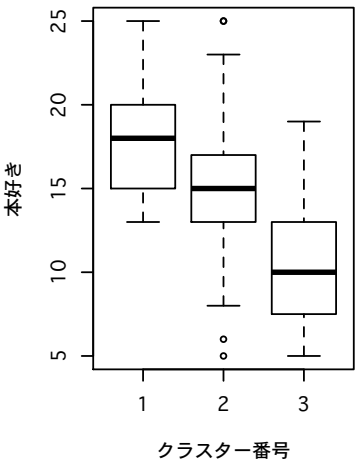

知識欲
クラスター番号


不可欠さ
クラスター番号


本好き
クラスター番号


知識欲
クラスター番号


不可欠さ
クラスター番号


本好き
クラスター番号