22 日目：コレスポンデンス分析

いろいろなクロス表を眺めていたら，ちょっとやってみたくなるのがコレスポンデンス分析ではないでしょうか（私だけ？）。対応分析ともよばれます。数量化 III 類のようなものといえばイメージできる人もいるかもしれませんね。

個人的事情を話せば，R に取り組んでみようと思ったのは，このコレスポンデンス分析をやってみたかったからです。

コレスポンデンス分析を平たくいえば，変数間，および変数のカテゴリ間の関係を簡潔に図示する方法のひとつといえるでしょう。似たもののまとまり具合を視覚化してくれます。

たとえば，以下のようなクロス表があったとします。これは「温泉」「テーマパーク」「リゾート」のそれぞれの旅行先について，どのような目的の時に適当だと思うか，いくつでも選んでもらった結果だとします。

	温泉	テーマパーク	リゾート
家族で	45	35	11
いっぱい遊ぶ	9	42	35
ゆっくり	28	23	25
豪華に	9	20	40
近場で	25	35	3
さわぐ	13	41	10
脱日常	20	25	10
リフレッシュ	33	12	12
のびのび	20	18	13

とりあえず，このデータをこの形式でエクセルに入力してください。そしてコピペで R に認識させます。やり方は昨日と同じです。１行目，１列目の行名，列名も読み込んでおきます。

```
x <- read.table(pipe("pbpaste"), header=TRUE, row.names=1,
fileEncoding="CP932")
```

コレスポンデンス分析の関数は R に２つあるようですが，MASS パッケージの corresp という関数を使ってみます。MASS がなければ，先にとってきておきましょう。

パッケージをよび出し，計算をさせて x.co に代入しておきます。そして，その中身を表示させてみます。

```
library(MASS)
x.co <- corresp(x, nf=2)
x.co
```

ちなみに corresp のカッコの中は，ファイル名と，nf=で抽出する軸の数を指定します。１つの図にまとめるなら，Ｘ軸とＹ軸という２軸なので２ですね。もちろん，常に２でよいかどうかは別問題ですが。

この結果は，以下のように返されます。

```
> library(MASS)
> x.co <- corresp(x, nf=2)
> x.co
First canonical correlation(s): 0.3759590 0.2555272

 Row scores:
                  [,1]       [,2]
家族で        -1.0329614 -0.3270235
いっぱい遊ぶ   1.2287221  0.7712943
ゆっくり       0.1985793 -0.8344208
豪華に         1.9086613 -0.7322165
近場で        -1.0936270  1.0302953
さわぐ        -0.1149727  1.8262362
脱日常        -0.4256637  0.3037221
リフレッシュ  -0.8666178 -1.6968603
のびのび      -0.1842117 -0.4785226

 Column scores:
                   [,1]       [,2]
温泉          -1.08393189 -0.9245510
テーマパーク  -0.09700997  1.1953393
リゾート       1.53021224 -0.7123953
>
```

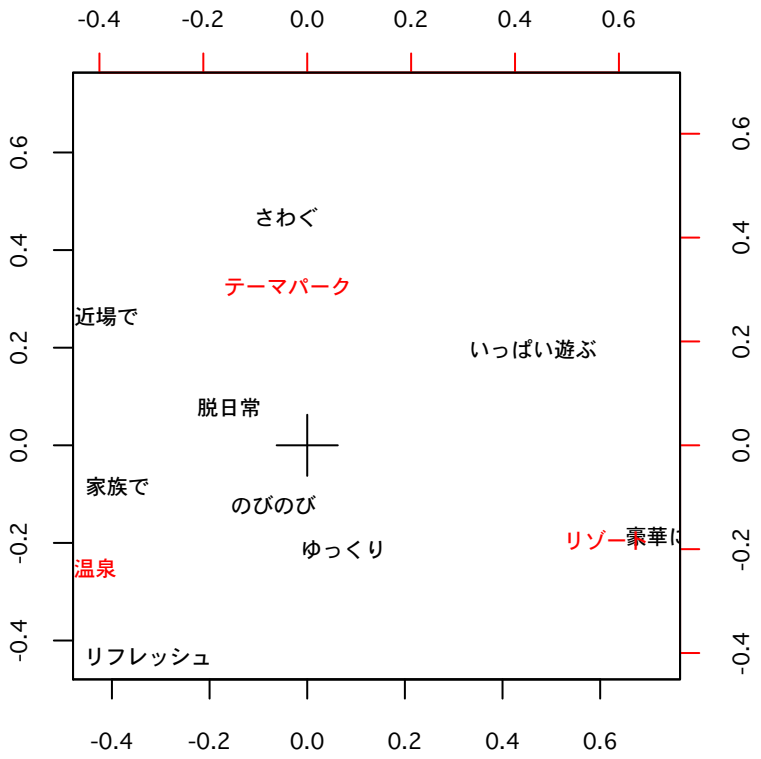
First canonical correlation(s)という部分は正順相関です。これを二乗したものが固有値になり，説明力がどの程度かを判断する材料になります。

次が行の得点，さらに列の得点となります。この得点を使って図示をします。

先に図を見た方がわかりやすいでしょうから，図示の命令をやってみます。par は図のコントロールです。日本語が入るので Osaka フォントを指定し，ps=で少し小さめのフォントサイズを指定しています（ここはお好みで）。そして biplot(x.co)と，カッコの中に先の計算結果をいれます。

```
par(family="Osaka", ps=8)
biplot(x.co)
```

これで，次のような図を出力してくれます。



これで行と列の変数およびカテゴリの位置関係を視覚的にとらえることができます。基本的に，近い位置にあるということは類似している，関係が強いというような意味になるので，たとえば，騒ぎたいならテーマパークを選びがち，豪華に過ごすならリゾート，家族でいくとか，リフレッシュ目的なら温泉，いっぱい遊ぶ場合は温泉というよりはテーマパークかリゾートを選びがち，のびのびとゆっくりは似たようなカテゴリ…などなどということがつかめるでしょう。データである表の結果と比較するとわかりやすいと思います。

さて，先に端折った固有値に関する件ですが，次のようにして計算できます。First canonical correlation(s)という部分は cor という名前で結果のオブジェクト(x.co)に入っています。それを取り出して２乗したものが固有値です。これを使ってさらに寄与率を計算します。これで第１軸，第２軸が，全体のどの程度を説明しているかがわかります。

```
x.eig <- x.co$cor^2
kiyoritu <- 100*x.eig / sum(x.eig)
kiyoritu
```

今回の結果だと，１軸が 68%ほど，２軸が 32%ほどとなります。今回は列が３列しかない小さな表なので，２軸の検討だけで十分でしょうが，もっと大きな表ならば，３軸以上の検討も必要です。

ちなみに，この軸（図）とデータの表がどのように対応するのかを確認するために，データを１軸の大きさで並べ替えてみましょう。

行も列も（カテゴリも変数も），上もしくは左に小さい値のものを，下もしくは右に行くほど値が大きなものになるように順番を入れ替えてみると以下のようになります。

	第1軸	温泉	テーマパーク	リゾート
近場で	−1.094	25	35	3
家族で	−1.033	45	35	11
リフレッシュ	−0.867	33	12	12
脱日常	−0.426	20	25	10
のびのび	−0.184	20	18	13
さわぐ	−0.115	13	41	10
ゆっくり	0.199	28	23	25
いっぱい遊ぶ	1.229	9	42	35
豪華に	1.909	9	20	40

把握しづらいかもしれませんが，右上から左下の対角線付近に，選択数が多いものが位置しています。コレスポンデンス分析は，こういう並べ替えを複数種類やっているようなものなのです。

表からだけではつかみにくい関係をつかむために，かなり使える方法ではないかと思います。