保健統計学実習

第5日目

第13回　主成分分析、因子分析、クラスター分析、コレスポンデンス分析
第14回 ┐ 解析実習まとめ：　統計解析ツール　R/R studio　解析例
第15回 ┘ （復習・課題の時間）

滋賀医科大学
NCD疫学研究センター　医療統計学部門

原田　亜紀子
（aharada@belle.shiga-med.ac.jp）

講義・演習スケジュール

1　R, EZRの使い方、データセットの読み込み、頻度集計、記述統計,相関 2　EZRのコード保存, R-studio　commander 3　エクセルの基礎(1)	8/31(木)
4　仮説検定の基礎,2群の比較(t検定, Wilcoxon検定) 5　カイ二乗検定、マクネマー検定 6　調査データ解析(1):調査票作成、データ入力	9/1(金)
7　重回帰分析 **8　ロジスティック回帰, 検査データの解析** 9　調査データ解析(2):(web調査ツールを使用した)調査票作成、データ入力	9/5(火)
10　分散分析 11　サンプルサイズ 12　調査データ解析(3):解析用データの作成	9/6(水)
13　主成分分析、因子分析、クラスター分析 14　解析実習・まとめ (復習・課題の時間)	9/8(金)

本日の実習内容

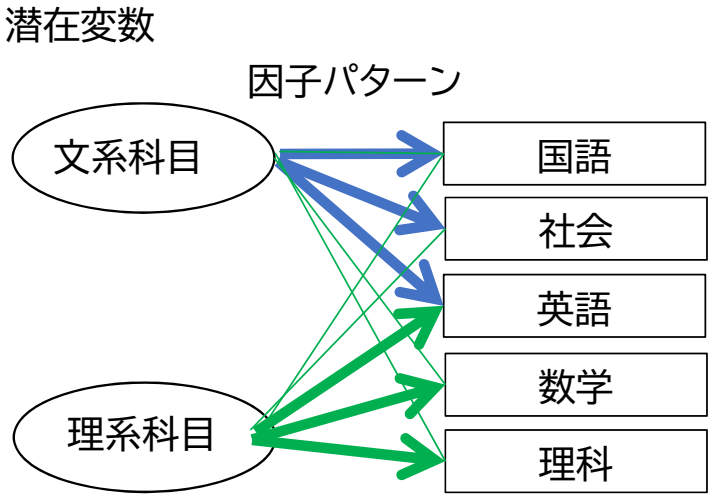
1. 主成分分析、因子分析、クラスター分析、コレスポンデンス分析

2. 実習・まとめ
   1）まとめ
   2）教科書など
   3）自宅環境で行うためには？　R解析例

第13回
1. 主成分分析、因子分析、クラスター分析、コレスポンデンス分析

主成分分析(Principal Component Analysis)・因子分析（Factor Analysis)

因子分析 （FA）



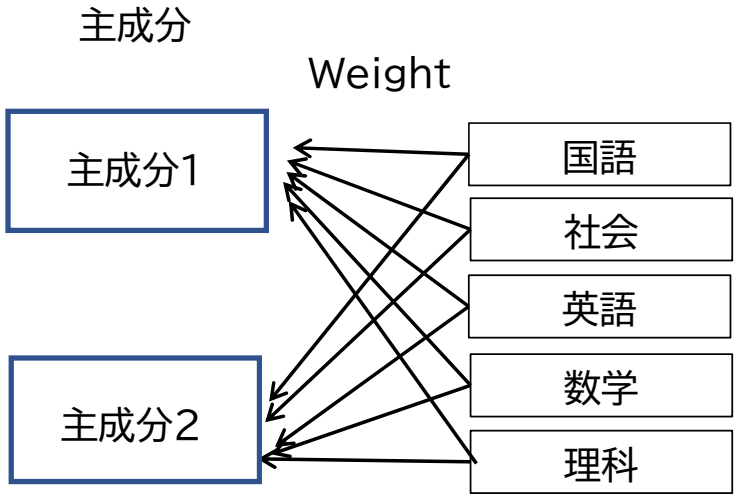
Unobservable ⟵ Observable

変数と要因の相関を説明する.

$$X_j = a_{j1}f_1 + a_{j2}f_2 + e_j$$

x: Observed variable, a: Factor loadings
f: Common factor, e: Uniqueness

主成分分析 （PCA）



Synthesis score ⟵ Observable

複合変数を、変数群の中で可能な限り最大の分散として計算する。

$$Z_j = a_{j1}X_1 + a_{j2}X_2 \quad * {a_{j1}}^2 + {a_{j2}}^2 = 1$$

z: Principal components, a: Principal component loadings
x: Observed variable

主成分分析

データの次元を下げる

- 主成分分析の目的は、測定された変数の集合から、元の変数の変動をできるだけ多くとらえる少数の独立した線形結合（主成分）を導き出すことである。
- 主成分分析は、次元削減の手法であると同時に、探索的データ分析ツールでもある。

主成分分析のアルゴリズム

①全データの重心(平均値)を算出する。

②重心から、データの分散が最大となる方向（第1主成分）を算出する。

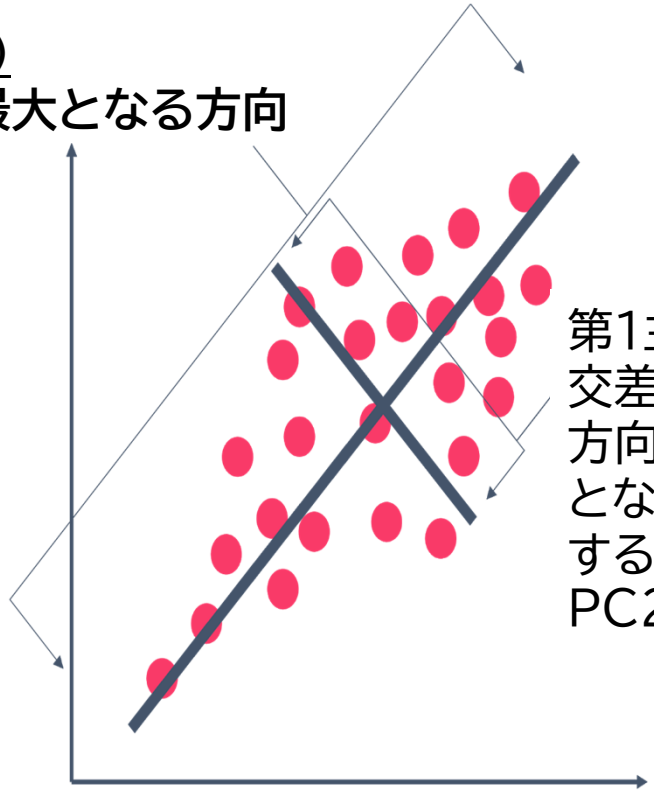
③第1主成分と直角（直交）に交わる方向で、分散が最大となる場所を算出する（第2主成分）。

④直近の主成分と直交する方向で分散が最大となる場所を算出する（第3主成分）。

⑤データの各次元について、手順④を繰り返す。

主成分分析

<u>第1主成分(PC1)</u>
データの分散が最大となる方向



PC1とPC2で
94%説明可能

【寄与率】
この主成分だけで元のデータの何パーセントを説明できるかを示す数値。

第1主成分と直角に交差する方向(直交方向)で分散が最大となる場所を計算する(第2主成分:PC2)。

ID1は理系科目が得意
ID2は両方得意

≪寄与率≫

	PC1	PC2	
寄与率	0.68	0.26	
累積寄与率	0.68	0.94	

≪負荷量≫

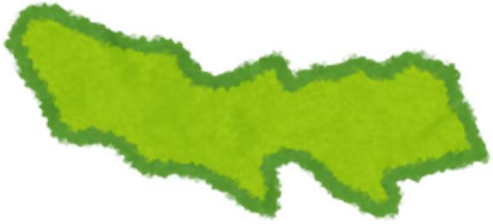
	PC1	PC2	
国語	0.08	0.93	
数学	0.96	-0.22	
理科	0.98	0.07	
社会	0.28	0.90	

≪主成分得点≫

ID	PC1	PC2	
1	47	-32	
2	51	28	

例題　東京都の自治体データ

- 東京都の自治体の各指標25指標
  - 市町村, 世帯あたり人数, 年齢15未満比率, 年齢65以上比率, 転入者_対人口比 転出者_対人口, 昼間人口比, 高齢単身世帯比率,
  - 第1次産業従業者数比, 第2次産業従業者数比, 第3次産業従業者数比,
  - 可住地面積比率, 耕地面積,_対可住面積比,
  - 課税所得_就業者1人あたり_千円, 小売業販売額_事業所あたり_百万円, 小売業販売額_売場面積あたり_万円, 国民健保一人あたり診療費_円, ごみリサイクル率_pct　千人あたり事業所数, 千人あたり幼稚園数,　千人あたり飲食店数, 千人あたり大型小売店数, 千人あたり病院数, 千人あたり老人ホーム数,千人あたり交通事故発生件数,千人あたり刑法犯認知件数
- 自治体に住みたいかどうかの調査結果を点数化した「人気度」
- **データセット:**
  - **TokyoSTAT_P25.xlsx**



主成分分析

主成分分析

```
#ライブラリの読込みlibrary(psych)
#主成分分析
resultPCA <- prcomp(DF[, -(1:3)], scale=TRUE)

#結果の要約
summary(resultPCA)
```

```
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     3.3971 2.1271 1.4791 1.27554 1.1
Proportion of Variance 0.4616 0.1810 0.0875 0.06508 0.0
Cumulative Proportion  0.4616 0.6426 0.7301 0.79518 0.8
                          PC11    PC12    PC13    PC14
Standard deviation     0.53764 0.45391 0.43334 0.36784
Proportion of Variance 0.01156 0.00824 0.00751 0.00541
Cumulative Proportion  0.96718 0.97542 0.98293 0.98835
                          PC20    PC21    PC22    PC23
Standard deviation     0.11608 0.10136 0.07726 0.06019
Proportion of Variance 0.00054 0.00041 0.00024 0.00014
Cumulative Proportion  0.99916 0.99957 0.99981 0.99996
```

- **固有値　Eigenvalue**
  - 最初の成分は、常に最も多くの分散を占め（したがって最も高い固有値を持つ）、次の成分は、できる限り残りの分散の多くを占め、そして以下同様とつづく。
- **累積　Cumulative**

  表は、
  - 固有値が1以上の成分で累積寄与率が約80%であることを示している
  - 上位5つの成分では60%程度

主成分分析： 数値の見方

- 固有値 Eigenvalue
  - 最初の成分は、常に最も多くの分散を占め（したがって最も高い固有値を持つ）、次の成分は、できる限り残りの分散の多くを占め、そして以下同様とつづく。

- 累積 Cumulative
  表は、
  - 固有値が1以上の成分で累積寄与率が約80%であることを示している
  - 上位5つの成分では60%程度

固有値

説明された分散の合計

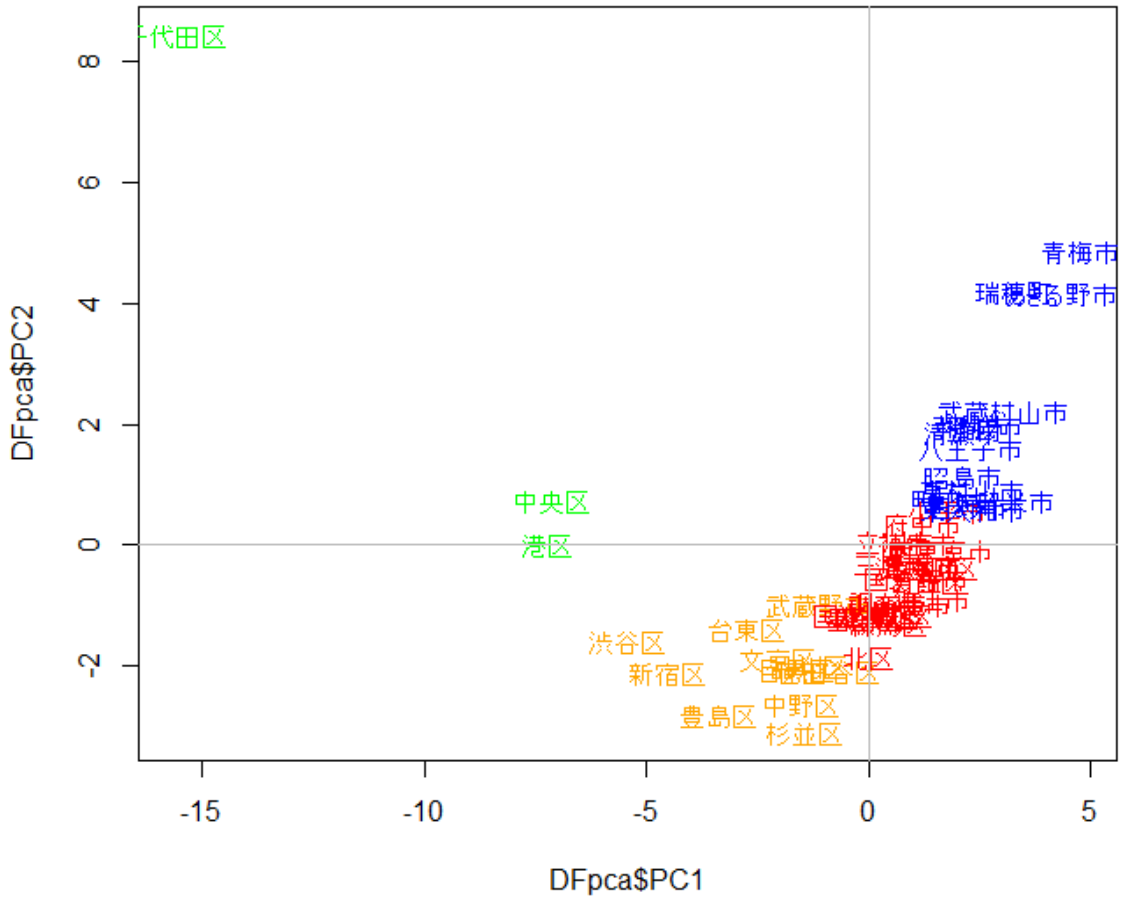
成分	初期の固有値 合計	初期の固有値 分散の %	初期の固有値 累積 %	抽出後の負荷量平方和 合計	抽出後の負荷量平方和 分散の %	抽出後の負荷量平方和 累積 %
1	7.037	28.150	28.150	7.037	28.150	28.150
2	2.867	11.469	39.619	2.867	11.469	39.619
3	1.853	7.413	47.031	1.853	7.413	47.031
4	1.719	6.878	53.909	1.719	6.878	53.909
5	1.571	6.284	60.193	1.571	6.284	60.193
6	1.519	6.077	66.270	1.519	6.077	66.270
7	1.216	4.863	71.133	1.216	4.863	71.133
8	1.033	4.134	75.267	1.033	4.134	75.267
9	1.011	4.043	79.310	1.011	4.043	79.310
10	.863	3.453	82.763			
11	.756	3.025	85.788			
12	.645	2.578	88.366			
13	.547	2.187	90.553			
14	.437	1.748	92.301			
15	.367	1.467	93.768			
16	.304	1.217	94.985			
17	.292	1.166	96.151			
18	.227	.909	97.060			
19	.188	.751	97.811			
20	.163	.654	98.465			
21	.157	.628	99.092			
22	.088	.353	99.445			
23	.071	.283	99.728			
24	.053	.211	99.939			
25	.015	.061	100.000			

因子抽出法: 主成分分析

主成分分析は回転を行っていないので抽出後の負荷量も同じ値

主成分分析：固有ベクトル（第3主成分まで）

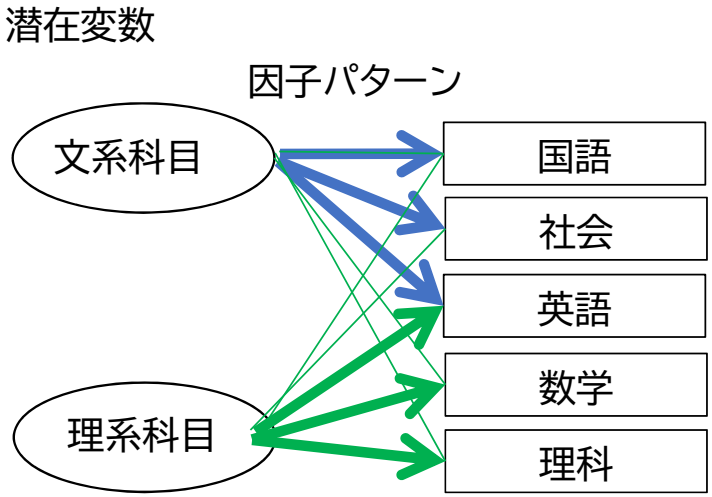
	PC1	PC2	PC3
世帯あたり人数	0.21817823	0.292007358	0.016652864
年齢15未満比率	0.17315873	0.282854276	0.107643674
年齢65以上比率	0.13827649	0.109660617	-0.462839305
転入者_対人口比	-0.28134219	-0.035917394	0.057305951
転出者_対人口比	-0.26794431	-0.117216398	0.117642903
昼間人口比_per	-0.24316331	0.243126759	-0.063282831
高齢単身世帯比率	-0.07399042	-0.218733757	-0.512410750
第1次産業従業者数比	0.14312036	0.166901626	0.250632303
第2次産業従業者数比	0.14036634	0.191454238	-0.128220532
第3次産業従業者数比	-0.14144537	-0.192633089	0.125130549
可住地面積比率	-0.11383968	-0.257891119	-0.087326397
耕地面積_対可住面積比	0.15650544	0.244937181	0.128000027
課税所得_就業者1人あたり_千円	-0.23531538	-0.003764779	0.128319792
小売業販売額_事業所あたり_百万円	-0.18853341	0.053126273	0.272612089
小売業販売額_売場面積あたり_万円m2	-0.23711598	-0.087377992	0.069927304
国民健保一人あたり診療費_円	0.13891741	0.086561773	-0.327457796
ごみリサイクル率_pct	0.14313231	0.127942834	0.261682729
千人あたり事業所数	-0.25378255	0.219289176	-0.056588990
千人あたり幼稚園数	-0.23505208	0.155132882	-0.083435715
千人あたり飲食店数	-0.26230625	0.191112884	-0.023531861
千人あたり大型小売店数	-0.25021531	0.226389573	0.003021881
千人あたり病院数	-0.15137829	0.284149431	-0.281049405
千人あたり老人ホーム数	0.12741348	0.292245219	0.020278499
千人あたり交通事故発生件数	-0.22590231	0.277312460	-0.014390255
千人あたり刑法犯認知件数	-0.24883963	0.187829606	-0.097897803


千代田区
中央区
港区
渋谷区
新宿区
豊島区
台東区
中野区
杉並区
北区
青梅市
八王子市
昭島市
DFpca$PC2
DFpca$PC1
8
6
4
2
0
-2
-15
-10
-5
0
5

因子分析

主成分分析･因子分析

因子分析



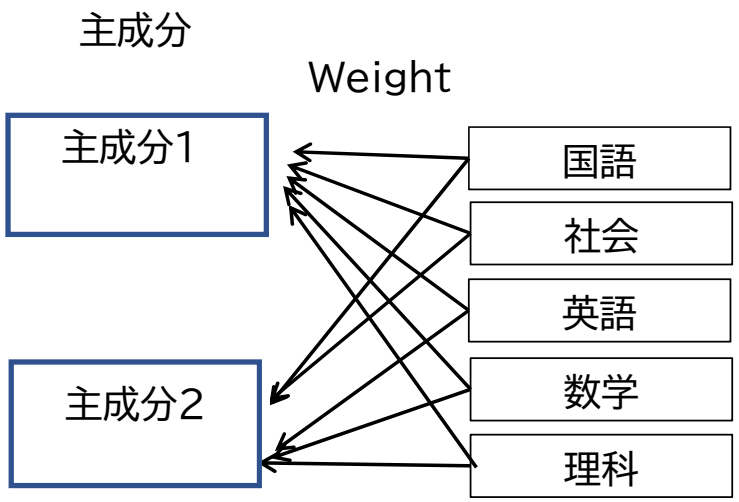
Unobservable ← Observable

変数と要因の相関を説明する.

$$X_j = a_{j1}f_1 + a_{j2}f_2 + e_j$$

x: Observed variable, a: Factor loadings
f: Common factor, e: Uniqueness

主成分分析



Synthesis score ← Observable

複合変数を、変数群の中で可能な限り最大の分散として計算する。

$$Z_j = a_{j1}X_1 + a_{j2}X_2 \quad * {a_{j1}}^2 + {a_{j2}}^2 = 1$$

z: Principal components, a: Principal component loadings
x: Observed variable

因子分析

- 因子分析は，観測された変数をより少ない数の（観測できない）潜在変数，または因子で記述しようとするものである．
  - 例では“文系科目”“理系科目”

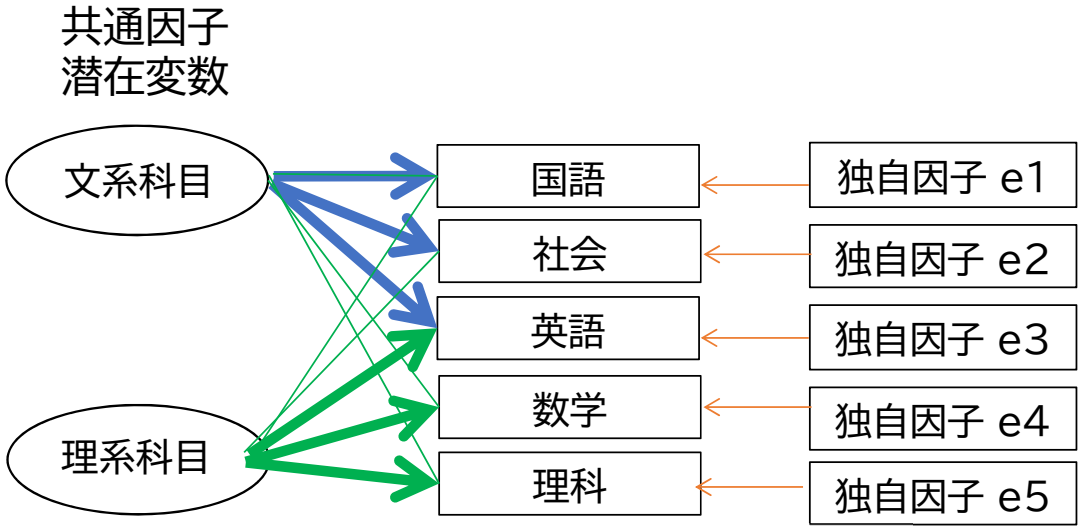
- 因子分析の目的は，

①観測された変数の意味のある解釈を，観測されない要因の観点から見つける

目に見えない概念

②変数の数を減らす

ことである。



因子分析のアルゴリズム

①固有値を算出する。
②因子負荷量の算出：共通因子の影響の強さを示す ”因子負荷量 ”を算出する。
【共通性】(Commonality)
・各観測変数がある因子群でどの程度説明できるかを示す数値。
・0(全く 説明できない)～1(完全に説明できる)の間の値である。
・1 - 共通性 = 固有の要因の量である。

【要因寄与度】
因子寄与度を観測変数の総数で割ることで、その因子が全体にどれだけ寄与(影響)しているかを見ることができる。

③ 因子軸を回転させる
各観測変数の因子負荷量を散布図グラフにプロットすると、共通因子がそのまま何を指しているのか分かりにくいことが多いので、解釈を容易にするために、各因子の数値が軸に沿うようにグラフの軸を回転させる。

④ 共通因子の名称を決定する。

⑤ 因子スコアの算出

回転

推定された因子負荷量行列を回転させ、単純な構造を得る（解釈しやすいようにする）

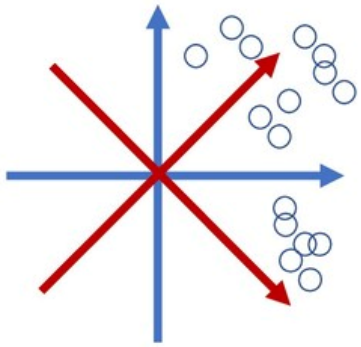
直交回転

各因子間の相関が0であること。
通常、バリマックス回転と呼ばれる方法が使われる。

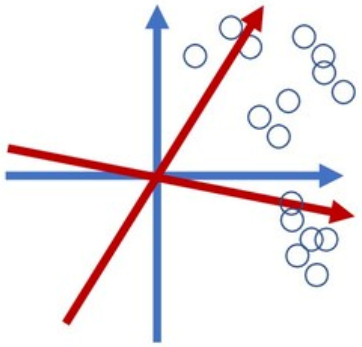
斜交回転

各因子の間に相関がある。
通常、プロマックス回転、直接オブリミンと呼ばれる方法などが使われる。

Varimax rotate



Promax rotate



因子抽出法

1. 主因子法(method=prinit)
   第1因子の因子寄与が最大となるように解が得られる（古典的方法）
   あまり用いられないが、ヘイウッドケースが生じにくい

2. 主成分法(method=principal) *default
   - 各因子の寄与率がなるべく等しくなるように解を求める。
   - **回転を伴わない主成分法の結果は、主成分分析の結果と同じになる**

3. 最尤法 (method=ml)
   - 解を確率密度により推定する。共分散構造解析でよく利用される。
   - 分布が歪んだデータでも正確な推定ができると言われている

Heywood（ヘイウッド）ケース：共通性の推定値が1より大きい場合

補足：因子抽出

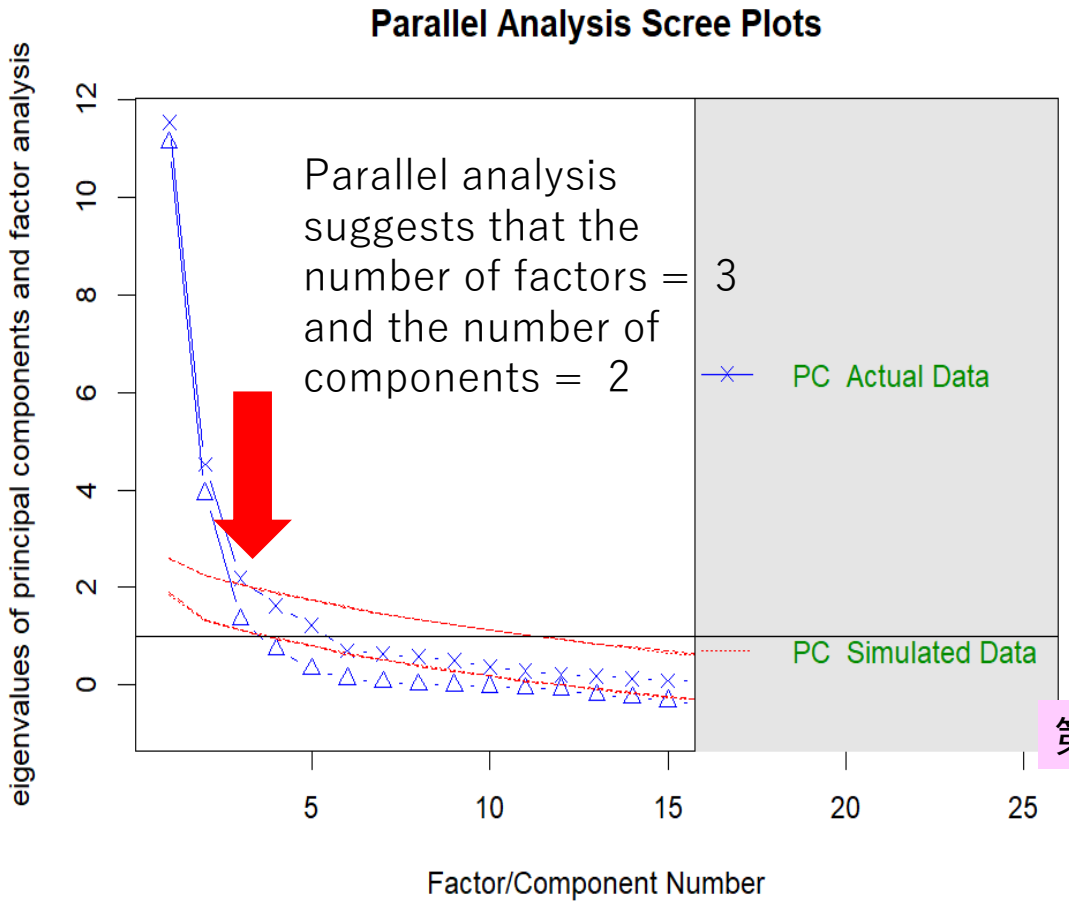
- 主因子法
  - 第1因子で説明される全体像を把握
- 最尤法
  - 全体像から傾向ごとに因子を作成してそれぞれを観測変数で説明
- 最小二乗法

＊最尤法と最小二乗法は誤差の重みづけが異なる
- 最小二乗法：すべての変数の誤差を同じ重み：共通性が低い項目の影響も強く受ける
- 最尤法：共通性が小さい項目は、重みを小さくして推定

＜方針＞
- **まずひとつの因子にまとめたい場合は「主因子法」→それ以外は「最尤法」で実施**
- **「最尤法」で実施したいが、各観測変数が正規分布をしていない場合は（対象数が少ない場合なども）、「一般化最小二乗法」、それでは不適解になってしまう場合は「重み付けのない最小二乗法」がよいか。**

Scree（スクリー） plot

```
result.prl <- fa.parallel(DF[, -(1:3)], fm="ml")
```



- Scree plotは、成分番号に対して固有値をグラフ化したもの。
- 第3成分から、線はほとんど平坦である。

→第何成分まで採択するかの目安

第3主成分ないしは第5主成分まで？

- resultFA <- fa(DF[, -(1:3)], nfactors=3,　　#因子数を指定
  fm = ”ml”,　　#pa 主因子法, ols 最小二乗法, ml 最尤法
  rotate = ”varimax”,　　#varimax 直交、promax 斜交
  scores = ”regression”)　#regression 回帰法

- resultFA <- fa(DF[, -(1:3)],
- nfactors=3,　　#因子数を指定
- fm = ”ml”,　#pa 主因子法, ols 最小二乗法, ml 最尤法
- rotate = ”varimax”,　#varimax 直交、promax 斜交
- scores = ”regression”)　#regression 回帰法#結果の表示#digits＝小数点以下表示桁の指定#sort=TRUEを指定(各項目ごとの因子負荷量がソートされる)
- print(resultFA, digits=2, sort=TRUE)

最尤法
バリマックス回転　例

Standardized loadings (pattern matrix) based upon correlation matrix

	item	ML1	ML2	ML3	h2	u2	com
千人あたり事業所数	18	0.97	0.16	0.14	1.00	0.0046	1.1
昼間人口比_per	6	0.97	0.11	0.12	0.97	0.0310	1.1
千人あたり大型小売店数	21	0.95	0.12	0.20	0.97	0.0327	1.1
千人あたり飲食店数	20	0.95	0.22	0.19	0.98	0.0171	1.2
千人あたり交通事故発生件数	24	0.95	0.00	0.14	0.92	0.0828	1.0
千人あたり刑法犯認知件数	25	0.91	0.23	0.05	0.88	0.1202	1.1
千人あたり幼稚園数	19	0.80	0.24	0.14	0.71	0.2866	1.2
千人あたり病院数	22	0.79	-0.01	-0.26	0.70	0.3041	1.2
転入者_対人口比	4	0.65	0.62	0.41	0.98	0.0240	2.7
課税所得_就業者1人あたり_千円	13	0.50	0.43	0.43	0.62	0.3763	2.9
小売業販売額_事業所あたり_百万円	14	0.49	0.17	0.45	0.47	0.5291	2.2
世帯あたり人数	1	-0.16	-0.92	-0.28	0.95	0.0502	1.3
年齢15未満比率	2	-0.06	-0.91	-0.09	0.83	0.1684	1.0
高齢単身世帯比率	7	0.03	0.74	-0.65	0.96	0.0372	2.0
転出者_対人口比	5	0.48	0.71	0.49	0.97	0.0283	2.6
耕地面積_対可住面積比	12	-0.13	-0.64	0.11	0.45	0.5544	1.1
千人あたり老人ホーム数	23	0.01	-0.63	0.16	0.43	0.5711	1.1
可住地面積比率	11	0.01	0.60	0.06	0.36	0.6362	1.0
小売業販売額_売場面積あたり_万円m2	15	0.44	0.58	0.34	0.65	0.3502	2.5
第3次産業従業者数比	10	0.09	0.54	0.23	0.35	0.6496	1.4
第1次産業従業者数比	8	-0.20	-0.54	0.08	0.34	0.6635	1.3
第2次産業従業者数比	9	-0.09	-0.54	-0.23	0.35	0.6536	1.4
ごみリサイクル率_pct	17	-0.24	-0.53	-0.11	0.36	0.6444	1.5
年齢65以上比率	3	-0.10	-0.19	-0.85	0.77	0.2337	1.1
国民健保一人あたり診療費_円	16	-0.15	-0.34	-0.48	0.36	0.6383	2.0

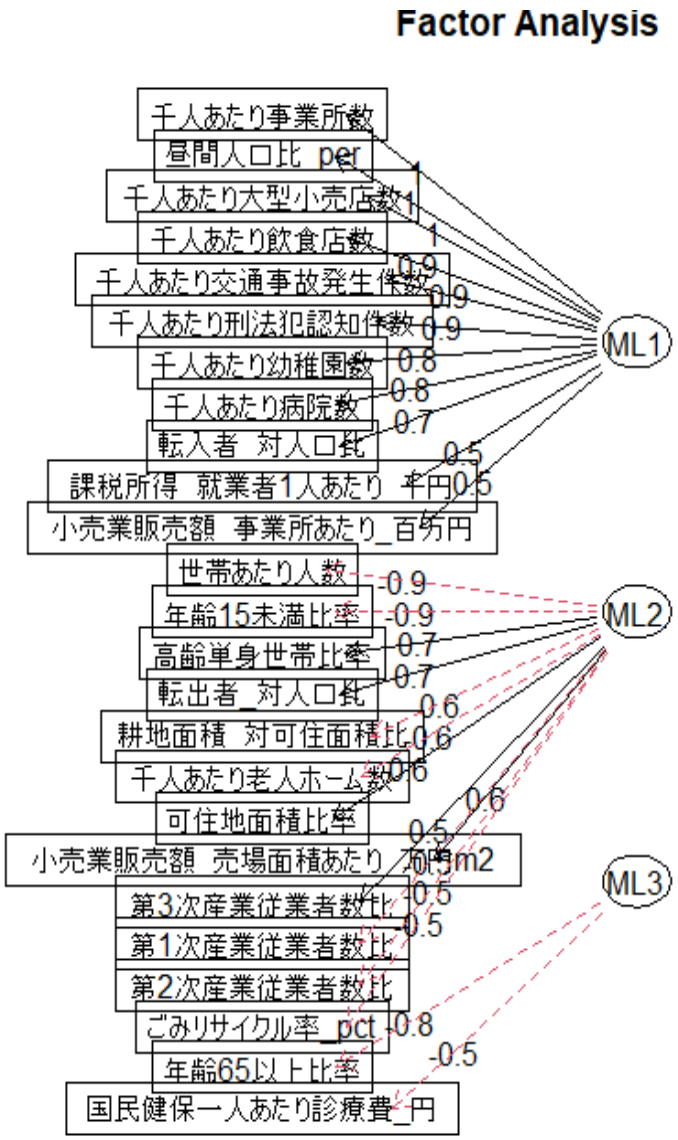
Standardized loadings (pattern matrix) based upon correlation matrix

	item	ML1	ML2	ML3	h2	u2	com
千人あたり事業所数	18	0.97	0.16	0.14	1.00	0.0046	1.1
昼間人口比_per	6	0.97	0.11	0.12	0.97	0.0310	1.1
千人あたり大型小売店数	21	0.95	0.12	0.20	0.97	0.0327	1.1
千人あたり飲食店数	20	0.95	0.22	0.19	0.98	0.0171	1.2
千人あたり交通事故発生件数	24	0.95	0.00	0.14	0.92	0.0828	1.0
千人あたり刑法犯認知件数	25	0.91	0.23	0.05	0.88	0.1202	1.1
千人あたり幼稚園数	19	0.80	0.24	0.14	0.71	0.2866	1.2
千人あたり病院数	22	0.79	-0.01	-0.26	0.70	0.3041	1.2
転入者_対人口比	4	0.65	0.62	0.41	0.98	0.0240	2.7
課税所得_就業者1人あたり_千円	13	0.50	0.43	0.43	0.62	0.3763	2.9
小売業販売額_事業所あたり_百万円	14	0.49	0.17	0.45	0.47	0.5291	2.2
世帯あたり人数	1	-0.16	-0.92	-0.28	0.95	0.0502	1.3
年齢15未満比率	2	-0.06	-0.91	-0.09	0.83	0.1684	1.0
高齢単身世帯比率	7	0.03	0.74	-0.65	0.96	0.0372	2.0
転出者_対人口比	5	0.48	0.71	0.49	0.97	0.0283	2.6
耕地面積_対可住面積比	12	-0.13	-0.64	-0.11	0.45	0.5544	1.1
千人あたり老人ホーム数	23	0.01	-0.63	0.16	0.43	0.5711	1.1
可住地面積比率	11	0.01	0.60	0.06	0.36	0.6362	1.0
小売業販売額_売場面積あたり_万円m2	15	0.44	0.58	0.34	0.65	0.3502	2.5
第3次産業従業者数比	10	0.09	0.54	0.23	0.35	0.6496	1.4
第1次産業従業者数比	8	-0.20	-0.54	0.08	0.34	0.6635	1.3
第2次産業従業者数比	9	-0.09	-0.54	-0.23	0.35	0.6536	1.4
ごみリサイクル率_pct	17	-0.24	-0.53	-0.11	0.36	0.6444	1.5
年齢65以上比率	3	-0.10	-0.19	-0.85	0.77	0.2337	1.1
国民健保一人あたり診療費_円	16	-0.15	-0.34	-0.48	0.36	0.6383	2.0

ML:各項目ごとの因子負荷量（各変数がどれだけ因子に寄与しているか）

h2:共通性--各変数の値の変動が因子でどれだけ説明できるかを表す

u2＝1-h2:独自性（uniqueness）--取りこぼしの度合（救えなかった情報）



ビジネス度

都会生活度

非高齢化度

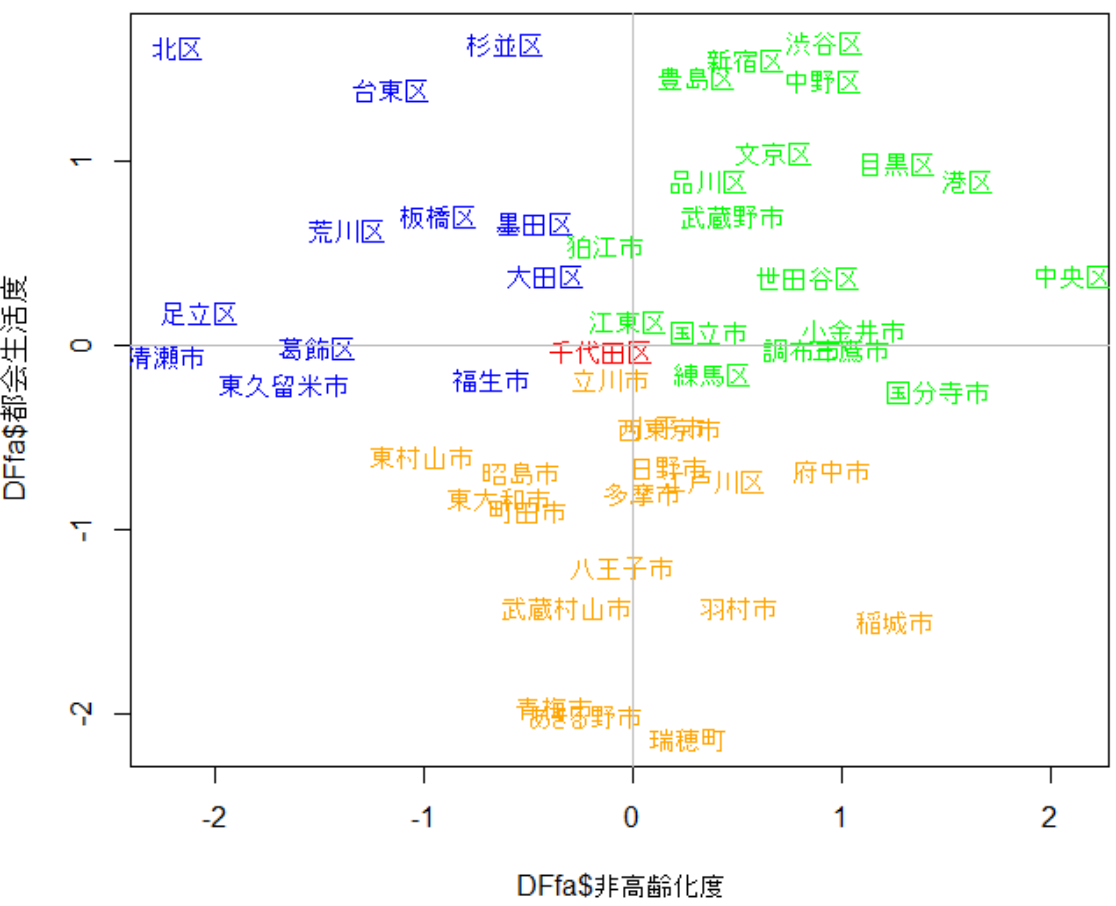
市町村

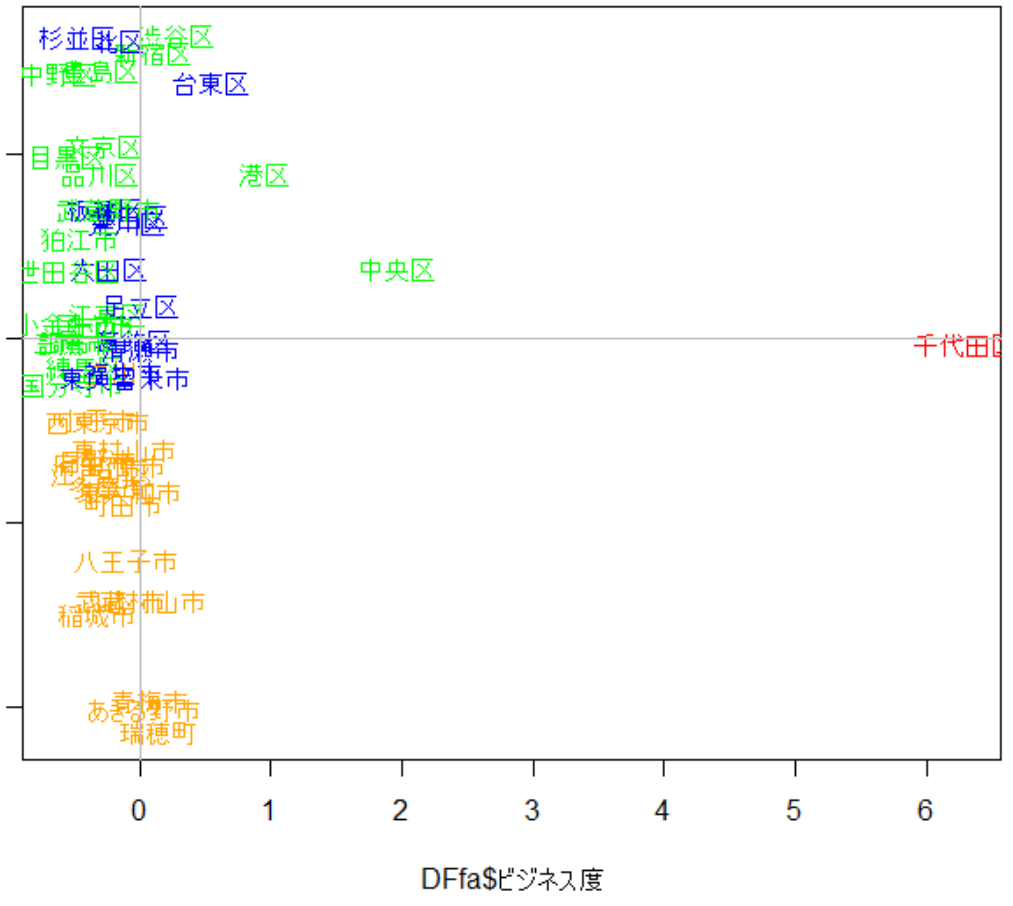
各市町村の因子得点

	ビジネス度	都会生活度	非高齢化度
千代田区	6.28952955	-0.027586855	-0.149562216
中央区	1.978500895	0.380486666	2.107701457
港区	0.959173252	0.896833027	1.609829713
新宿区	0.108925127	1.561911547	0.549955397
文京区	-0.259646529	1.054773559	0.679235552
台東区	0.559442656	1.394625835	-1.152476274
墨田区	-0.061681736	0.673131427	-0.456751709
江東区	-0.240627283	0.13066321	-0.0121625
品川区	-0.288504431	0.898379753	0.374448283
目黒区	-0.543939983	0.996072929	1.272587375
大田区	-0.215838833	0.383747425	-0.408010983
世田谷区	-0.540403354	0.37097304	0.83914382
渋谷区	0.281888713	1.652728466	0.928721734
中野区	-0.611870679	1.442986897	0.915149195
杉並区	-0.462077032	1.646218736	-0.604875212
豊島区	-0.289235033	1.464530994	0.31288243
北区	-0.158709109	1.623151525	-2.171938947
荒川区	-0.087915345	0.632385909	-1.358090375
板橋区	-0.253579023	0.708551199	-0.925024645
練馬区	-0.403807277	-0.151484721	0.388270014
足立区	0.022100068	0.180164964	-2.067885827
葛飾区	-0.029866103	-0.011593432	-1.505881164
江戸川区	-0.279638467	-0.735431315	0.390862616

	ビジネス度	都会生活度	非高齢化度
八王子市	-0.099671387	-1.199477087	-0.04678258
立川市	-0.121018774	-0.177891696	-0.103027833
武蔵野市	-0.232449748	0.70721736	0.480050928
三鷹市	-0.502386687	-0.013368301	1.050097454
青梅市	0.089412329	-1.96221811	-0.36424875
府中市	-0.355487881	-0.681085313	0.963194708
昭島市	-0.088488637	-0.685701167	-0.523373265
調布市	-0.457058102	-0.022606221	0.813580803
町田市	-0.114558365	-0.896725464	-0.497786839
小金井市	-0.553205122	0.087822536	1.058275538
小平市	-0.319514093	-0.441150658	0.170051147
日野市	-0.308764705	-0.655460556	0.176919321
東村山市	-0.117527413	-0.604472493	-1.005092723
国分寺市	-0.516158616	-0.249588777	1.462694427
国立市	-0.325528804	0.076632321	0.375560108
福生市	-0.117075838	-0.182515334	-0.673841624
狛江市	-0.447201925	0.549907778	-0.127769793
東大和市	-0.069875086	-0.830438297	-0.634392094
清瀬市	0.014032578	-0.055480602	-2.226392395
東久留米市	-0.096283925	-0.204713	-1.662185014
武蔵村山市	0.017797519	-1.4172516	-0.310523763
多摩市	-0.233301151	-0.799006611	0.052659121
稲城市	-0.318010895	-1.496100465	1.265650401
羽村市	-0.082247289	-1.4186142	0.514181113
あきる野市	0.045441875	-2.010095498	-0.220648725
西東京市	-0.322066024	-0.448273517	0.182568193
瑞穂町	0.158976123	-2.135565814	0.274454402

因子得点でプロット


北区
杉並区
台東区
豊島区
新宿区
渋谷区
中野区
文京区
品川区
目黒区
港区
荒川区
板橋区
墨田区
武蔵野市
狛江市
大田区
世田谷区
中央区
足立区
江東区
国立市
小金井市
清瀬市
葛飾区
千代田区
調布市
三鷹市
東久留米市
福生市
立川市
練馬区
国分寺市
西東京市
東村山市
昭島市
日野市
府中市
東大和市
多摩市
町田市
八王子市
武蔵村山市
羽村市
稲城市
瑞穂町
DFfa$都会生活度
DFfa$非高齢化度
-2
-1
0
1
2


杉並区
北区
渋谷区
新宿区
中野区
豊島区
台東区
文京区
目黒区
品川区
港区
狛江市
世田谷区
中央区
足立区
千代田区
東村山市
町田市
八王子市
稲城市
瑞穂町
DFfa$ビジネス度
0
1
2
3
4
5
6

因子分析ポイント

- 因子抽出法、回転方法など様々な組み合わせが存在
  - ヘイウッドケース
    →変数の除去（相関が高い変数）
    →単一変数で構成されている因子がないか
    →主因子法の選択
    →別な回転の選択

  - 因子負荷量の小さい変数の対応

  - 相関の強さ：変数の選択、直交、**斜交回転の選択**

  →［再解析の必要性］ 様々なパターンを試す必要がある（ある意味正解はない）

クラスター分析

変数、対象者を分類する

クラスター分析

教師なし分類(クラスタリング)

＜変数データの分類＞

- 変数のクラスタリング

  数値変数の集合を不連続あるいは階層的なクラスタに分割する

＜ケースデータの分類＞

- k-means法による個体のクラスタリング　・・・あらかじめ分類するクラスター数を決定

  1つ以上の量的変数から計算された**距離に基づいて**、不連続なクラスター分析を行う

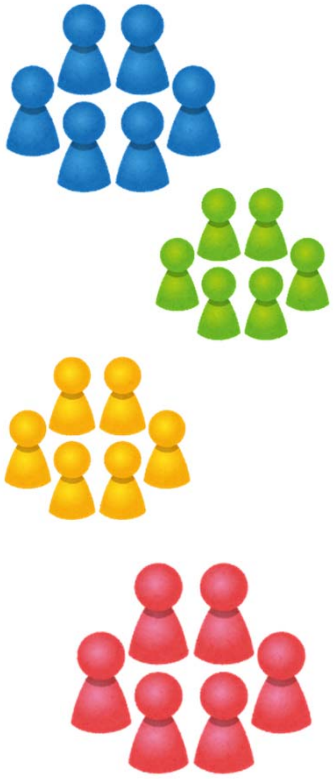
- CLUSTER **距離に基づく**階層的クラスタリング

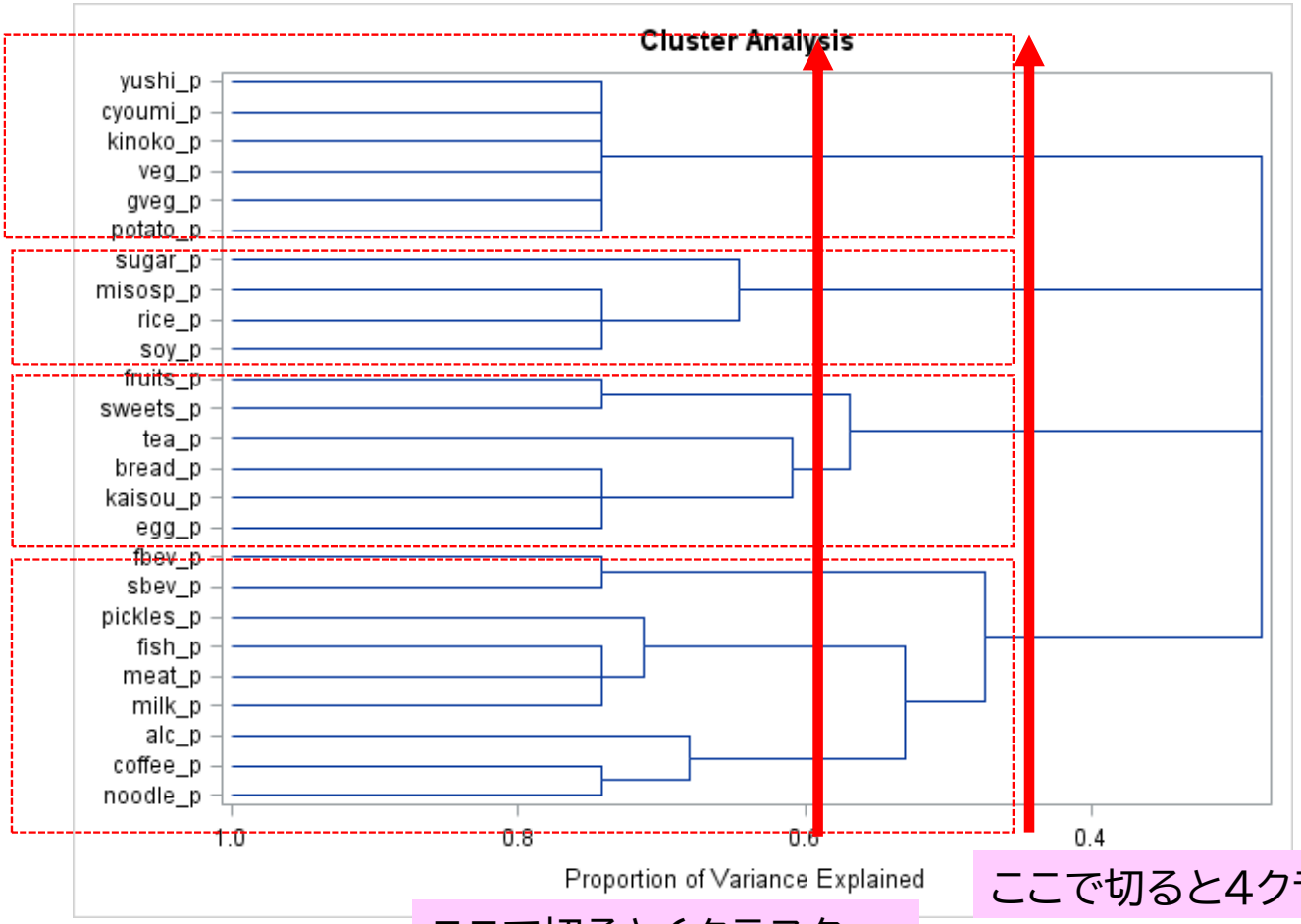
  単一連結 , 完全連結 , 平均連結 , ウォード, セントロイド, 密度

距離の決め方　いろいろあります

1. **Ward法:**「各クラスターに属するケース の平均値を出し, その平均値から各ケースの差を求め, 差を2乗したうえで, 全クラスターを合算する」(平方和 指標 E)ものである。この値が最も低いものを融合の対象とする
2. **グループ間平均連結法 :**ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)のそれぞれからひとつのケースを選択してできる組み合わせ(AD, AE, AF, BD, BE, BF, CD, CE, CF)の距離を平均し、その値を両クラスター間の距離であるとする。
3. **グループ内平均連結法 :**ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)に属するすべてのケースから作る可能性のある2ケースの組み合わせ (AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, CF, DE, DF, EF)の距離を 平均し、その値を両クラスター間の距離であるとする。
4. **最近隣法 :**ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)のそれぞれからひとつのケースを選択してできる組み合わせのすべて(A D, AE, AF, BD, BE, BF, CD, CE, CF)の距離のうち最も短いものをもって, 両クラスター間 の距離であるとする。
5. **最遠隣法 :**ひとつのクラスター(ケース:A, B, C)ともうひとつのクラスター(ケース:D, E, F)のそれぞれからひとつのケースを選択してできる組み合わせのすべて (AD, AE, AF, BD, BE, BF, CD, CE, CF)の距離のうち最も遠いものをもって, 両クラスタ ー間の距離であるとする。
6. **重心法 :**ひとつのクラスターについて, ケース間の距離の測定に用いる複数の変数の平均でク ラスターの座標を求め, これをそのクラスターの重心とする。クラスターを構成するケース数で重み付けを行ったうえでクラスターの重心間の距離を求め, これが最も短いクラスター群を融合させる。

“変数”のクラスタリング　食事調査の例





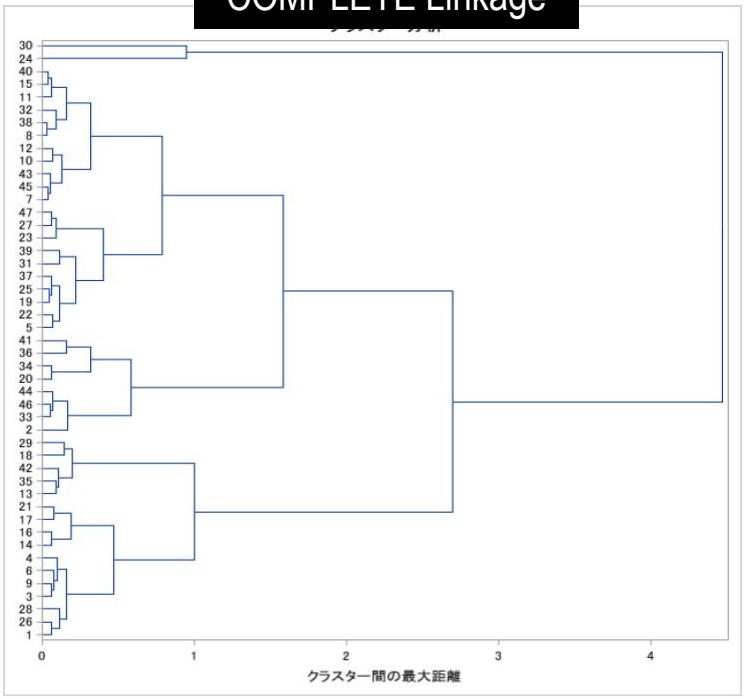
ここで切ると4クラスター

ここで切ると6クラスター

クラスター分析　階層的　対象者を分類

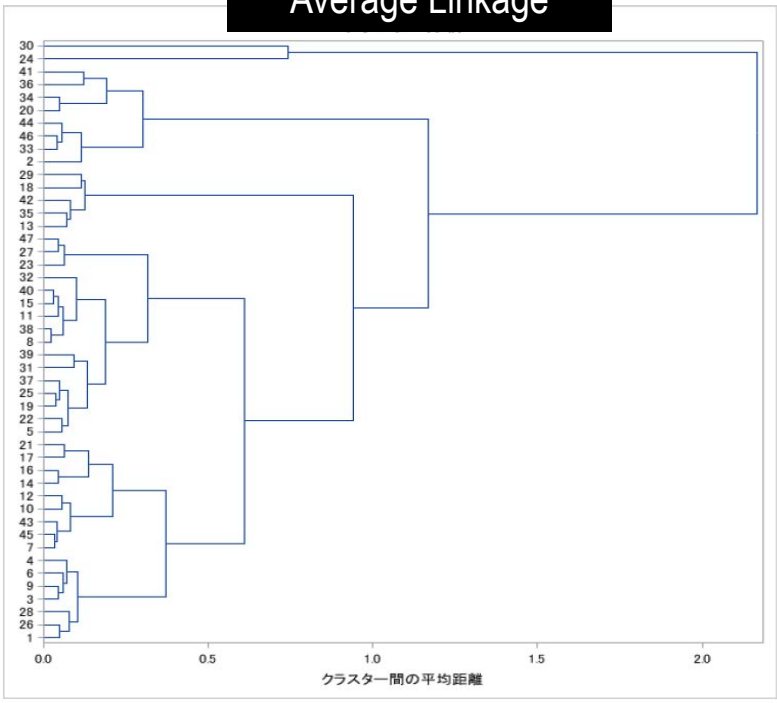
数字は対象者

COMPLETE Linkage



2つのクラスタ間の距離は、1つのクラスタの観測値と他のクラスタの観測値の間の最大距離である。

Average Linkage



2つのクラスタ間の距離は，各クラスタに1つずつあるオブザベーションのペアの間の平均距離である．

例:チョコレート・キャンディーデータ

データセット
Candy_Bars.csv
Candy_Bars.sav

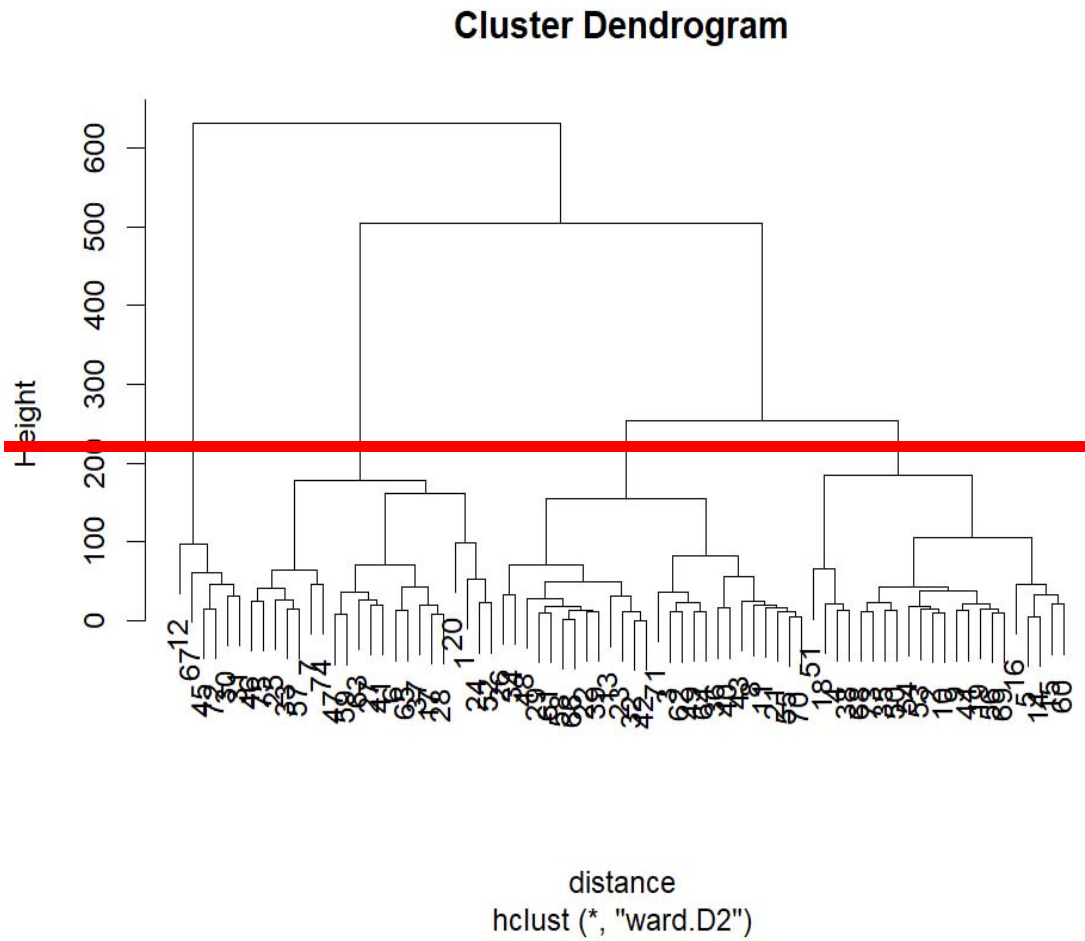
各社のチョコレートキャンディ(75銘柄)の栄養素含有量のデータから、その特徴によりいくつかのクラスターに分けたい

- 階層的クラスター分析
- K-means法

①階層クラスタリング　Ward法

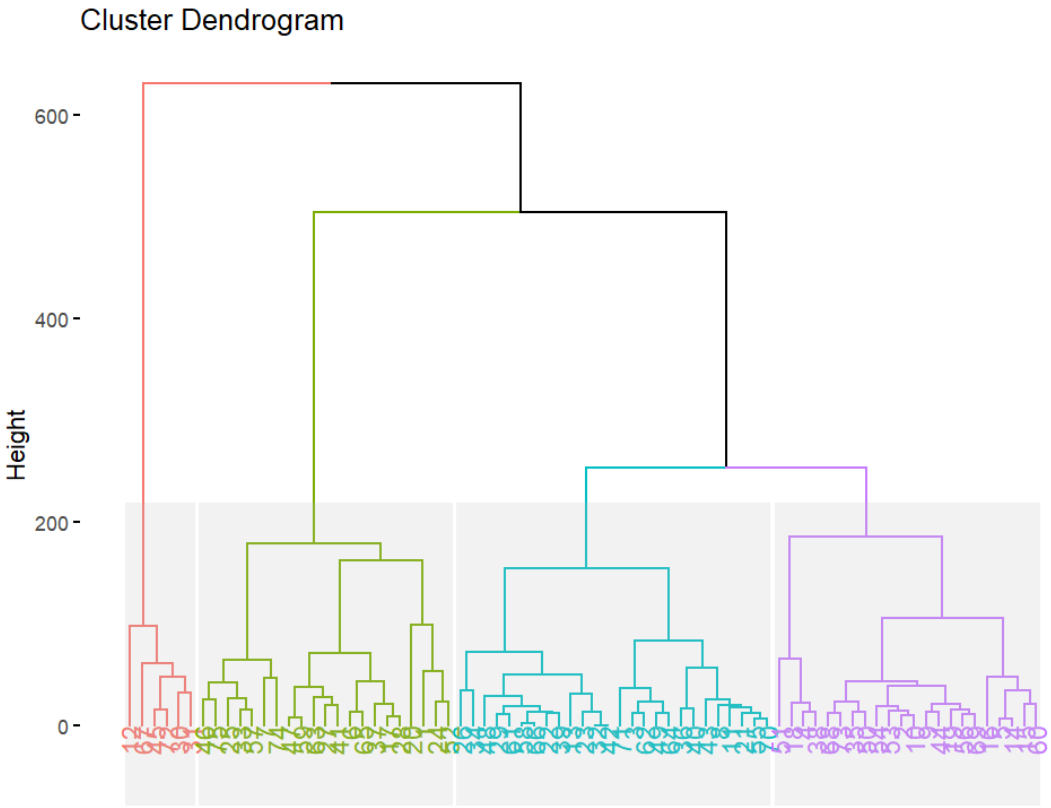
```
# 非類似度(距離)を計算

Candy_Bars <- read.csv("F:/■■■大学/◎◎滋賀医大/■■講義関連/京都府大/講義資料/2023_R/##第5回/dataset/Candy_Bars.csv")


data <- Candy_Bars[,4:11]

distance <- dist(data)

#ユークリッド距離を求める# 樹形図作成
hc <- hclust(distance, "ward.D2")

#プロット
plot(hc)res <- cutree(hc, k =3)
```

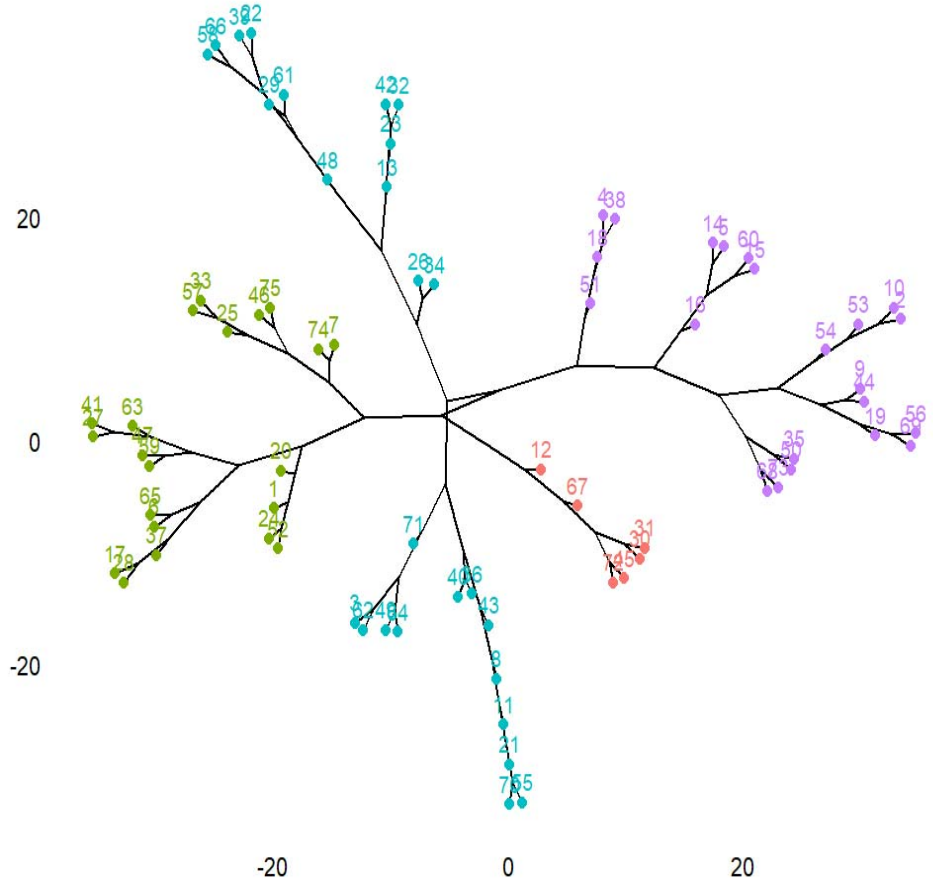
ブランド名	名前	サービング/pkg	オンス/pkg	カロリー	総脂肪(g)	飽和脂肪(g)	コレステロール(g)	塩分(mg)	炭水化物(g)	食物繊維(g)	糖分(g)	タンパク質(g)
M&M/Mars	Snickers Peanut Butter	1	2	310	20	7	5	150	28	1	23	6
Hershey	Cookies 'n' Mint	1	1.55	230	12	6	10	80	27	1	21	4
Hershey	Cadbury Dairy Milk	3.5	5	220	12	8	10	45	24	1	21	3
M&M/Mars	Snickers	3	3.7	170	8	3	5	85	21	1	17	3
Charms	Sugar Daddy	1	1.7	200	2.5	2.5	2	100	43	0	28	1
M&M/Mars	Twix Peanut Butter	1	1.71	260	16	5	5	130	26	2	17	5
Hershey	Twizzler	1	2.2	190	1.5	0	0	150	42	0	24	2
Tobler	Toblerone	1	1.23	190	11	7	5	25	21	0	19	2
Nestle	Crunch	1	1.55	230	12	7	5	60	28	1	23	3
Hershey	Almond Joy	2	3.22	230	13	8	2	85	25	3	17	2
Sherwood	Elana Mint	1	1.6	200	10	6	15	10	29	2	26	2
Hershey	Krackel	1	2.6	390	21	13	10	110	45	1	35	5



	1	2	3	4
100 Grand	0	1	0	0
3 Musketeers	1	0	0	0
5th Avenue	1	0	0	0
Abba-Zabba	1	0	0	0
Almond Joy	0	1	0	0
Almond Roca	1	0	0	0
Baby Ruth	1	0	0	0
Bar None	0	0	1	0
Big Cherry	0	0	1	0
Big Hunk	1	0	0	0
Bit-O-Honey	0	1	0	0
Butterfinger	1	0	0	0
Toblerone	0	0	1	0
Twix Caramel	1	0	0	0
Twix Peanut Butter	1	0	0	0
Twizzler	1	0	0	0
U-No (Blue)	0	0	1	0
U-No (Green)	0	0	1	0
Whatchamacallit	1	0	0	0
Whoppers	1	0	0	0
York Peppermint Patty	0	0	1	0

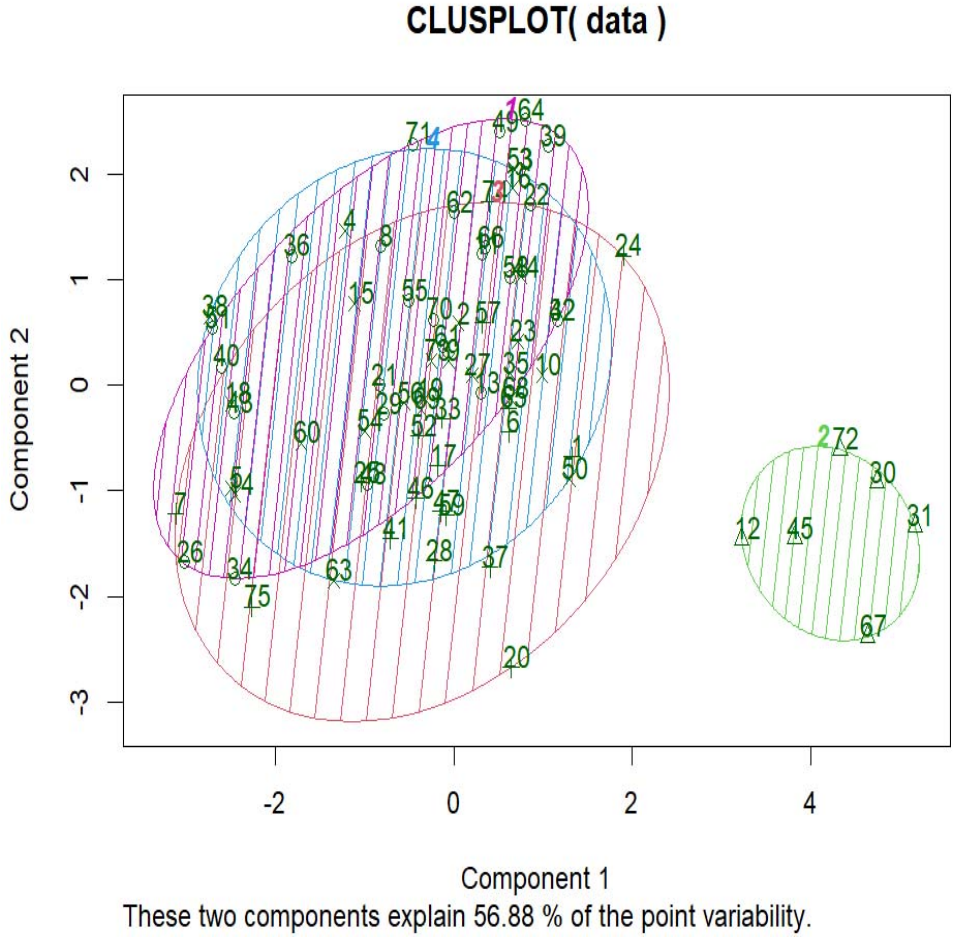
##別
library(tidyverse)hc %>% factoextra::fviz_dend(k=4, rect=TRUE, rect_fill=TRUE)library(igraph)hc %>% factoextra::fviz_dend(k=4, rect=TRUE, rect_fill=TRUE, type="phylogenic")

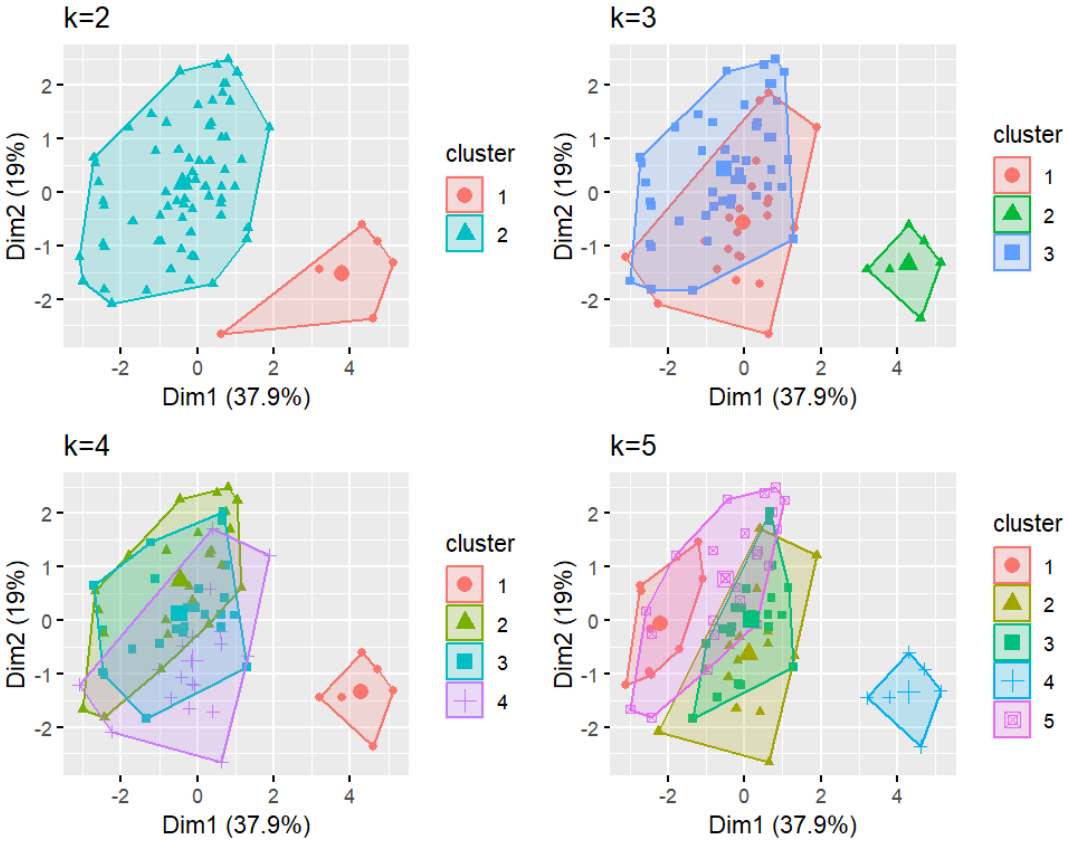


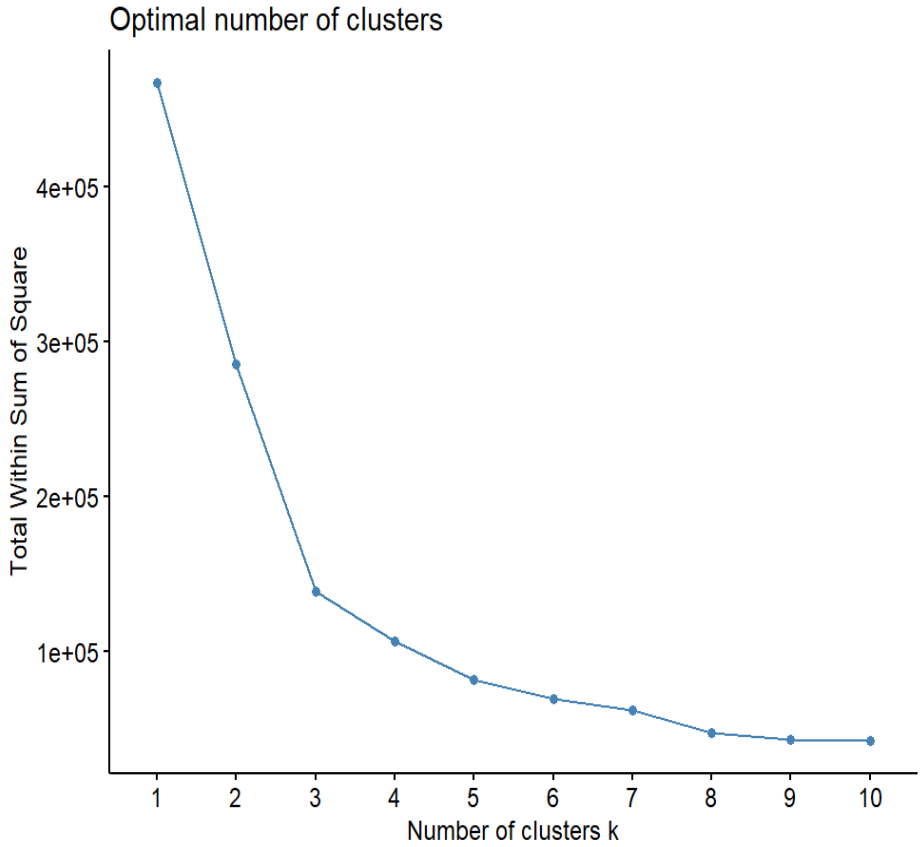


②k-means法による

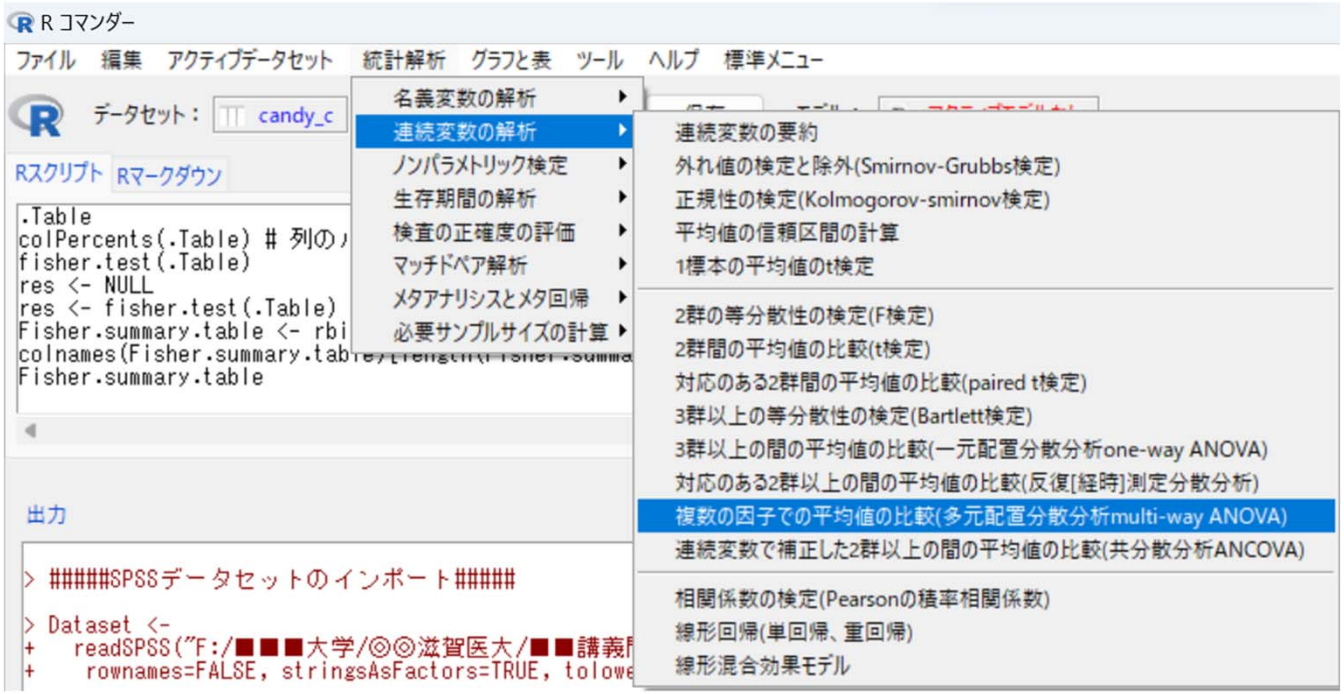
```
##k-means

km<-kmeans(data,4)
result2 <- km$cluster
write.csv(result2,"result2.csv")

#plot
library(cluster)
clusplot(data, km$cluster, color=TRUE,
shade=TRUE, labels=2, lines=0)

#data出力
table2 <- table(answer, result2)
write.csv(table2,"table2.csv")
```




k=2
k=3
k=4
k=5
Dim1 (37.9%)
Dim2 (19%)
cluster
1
2
3
4
5


Optimal number of clusters
Total Within Sum of Square
Number of clusters k
1e+05
2e+05
3e+05
4e+05

階層（Ward法）とk-means法の比較



「Candy_cluster.csv 」に
・階層
・k-meansのクラスター情報
・チョコバーの属性
がまとめてある。

階層、k-meansのクラスターごとの特徴を比較してみる

→ ANOVA
箱ひげ図

①Wards法による分類　4つで指定

		カロリー		コレステロール.g.		塩分.mg.		炭水化物.g.		糖分.g.	
n		mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
1	21	258.7	33.2	4.5	6.1	136.4	32.1	34.1	8.9	25.0	7.4
2	22	212.5	33.4	4.3	5.1	82.3	14.5	27.5	6.7	20.4	4.8
3	26	216.5	25.0	5.8	5.1	30.6	14.6	28.8	9.5	23.3	8.3
4	6	415.0	19.7	9.2	4.9	52.5	34.9	41.0	5.4	31.7	2.

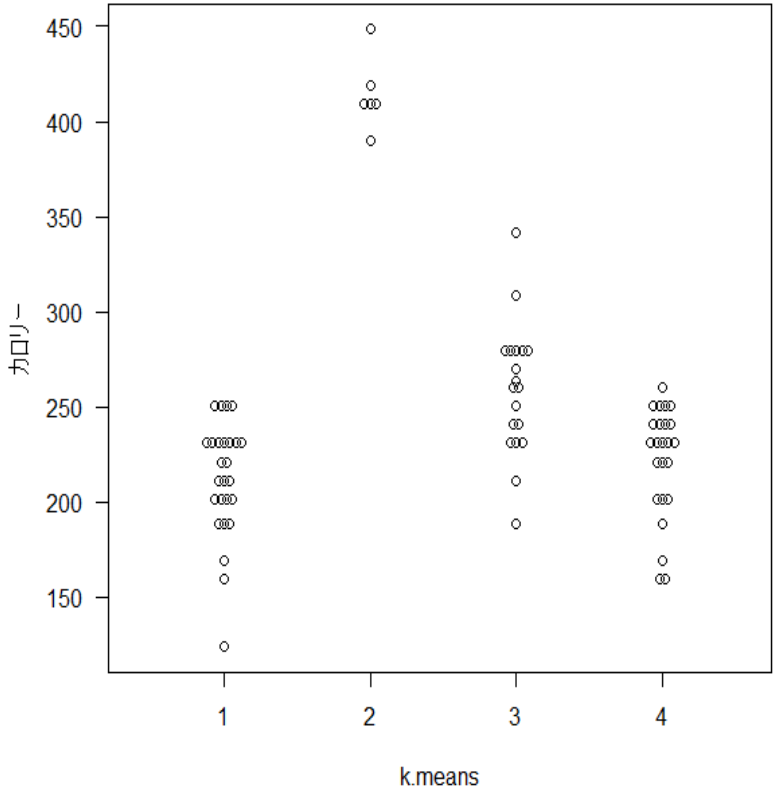
②K-means法による分類　4つで指定

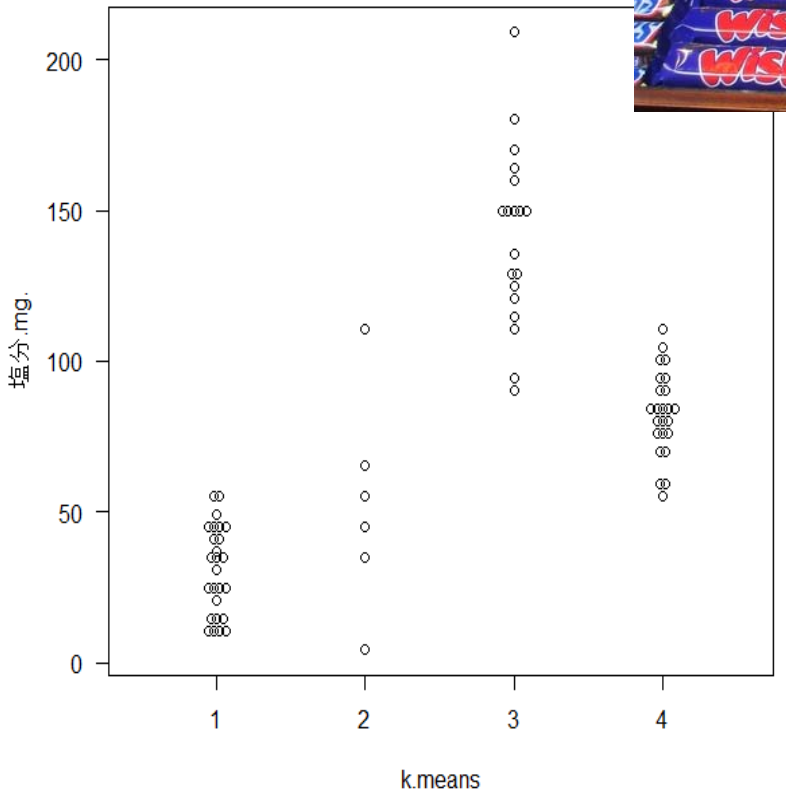
4つのクラスター

		カロリー		コレステロール.g.		塩分.mg.		炭水化物.g.		糖分.g.	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	26	211.73	29.83	5.65	5.21	30.62	14.60	28.65	9.57	23.08	8.58
2	6	415.00	19.75	9.17	4.92	52.50	34.89	41.00	5.44	31.67	2.80
3	19	259.05	34.95	4.42	6.18	141.26	29.58	33.37	8.80	24.79	7.67
4	24	221.25	28.79	4.58	4.99	82.92	14.14	28.88	7.59	21.17	4.60

②k-means法による

4つのクラスター





分類方法による一致度

		階層			
		1	2	3	4
k.means	1	0	1	25	0
	2	0	0	0	6
	3	19	0	0	0
	4	2	21	1	0

①低塩分、低糖質
②カロリー、糖質
③塩分
④普通

チョコバーの分類(k-meansで)

1		2		3		4	
36	Adams &	12	Hershey	75	Annabelle	63	Annabelle
32	Annabelle	30	Hershey	24	Brown &	60	Annabelle
42	Annabelle	31	Hershey	7	Hershey	14	Bit-O-Honey
3	Hershey	45	Hershey	17	Hershey	5	Charms
40	Hershey	67	Hershey	33	Hershey	27	Hershey
49	Hershey	72	Hershey	47	Hershey	2	Hershey
58	Hershey			25	Leaf	10	Hershey
62	Hershey			46	Leaf	16	Hershey
64	Hershey			74	Leaf	35	Hershey
66	Hershey			1	M&M/Mars	44	Hershey
71	Hershey			6	M&M/Mars	50	Hershey
26	Just Born			41	M&M/Mars	53	Hershey
13	M&M/Ma			57	M&M/Mars	69	Hershey
34	M&M/Ma			59	M&M/Mars	73	Hershey
39	M&M/Ma			65	M&M/Mars	23	Hershey
55	M&M/Ma			28	Nestle	4	M&M/Mars
61	M&M/Ma			37	Nestle	18	M&M/Mars
29	Myerson			20	Pearson	19	M&M/Mars
21	Nestle			52	Standard	54	M&M/Mars
70	Nestle					68	M&M/Mars
11	Sherwood					56	Nabisco
22	Sherwood					9	Nestle
8	Tobler					15	Nestle
43	Tootsie					38	Weider
48	Tootsie						
51	Weider						

Cup O Gold	
U-No (Green)	
U-No (Blue)	
Cadbury Dairy Milk	
York Peppermint	
Kisses Almond	
Milk Chocolate	
Cadbury Fruit &	
Kisses	
Symphony (Red)	
Cadbury Caramello	
Super Hot Tamales	
M&Ms Peanut	
Skittles	
Dove	
M&Ms Almond	
M&Ms Plain	
Big Cherry	
Raisinet	
Chunky	
Elana Mint	
Elana Mocca	
Toblerone	
Jr Mints	
Charleston Chew	
Tiger Sport	

Krackel	
Mr. Goodbar	
Golden Collection	
KitKat	
Special Dark	
Milk Chocolate	

Big Hunk	
Almond Roca	
Twizzler	
Twix Caramel	
Reese's Peanut	
5th Avenue	
Payday	
Whoppers	
Heath	
Snickers Peanut	
Twix Peanut Butter	
3 Musketeers	
Snickers Munch	
Milky Way	
Whatchamacallit	
Butterfinger	
Baby Ruth	
Peanut Nut Roll	
Peanut Butter	

Abba-Zabba	
Look!	
Bit-O-Honey	
Sugar Daddy	
Rolo	
Cookies 'n' Mint	
Almond Joy	
Skor	
Reese's Peanut Butter	
Symphony (Blue)	
Mound	
Cadbury Roast Almond	
Reese's Pieces	
Reese's Nutrageous	
Bar None	
Snickers	
Milky Way Lite	
Mars	
Milky Way Dark	
M&Ms Peanut Butter	
Planters Original Peanut	
Crunch	
100 Grand	
Tiger Milk	

因子分析、主成分分析の例題

1. クラスター分析
因子得点(ビジネス度、都会生活度、非高齢化度)
それぞれ用い、クラスター分析を行う。
各クラスターにおいて、特徴と所属する市町村を確認する

2. クラスターごとの市町村人気度スコアを比較する

＃＃因子得点→クラスター分析

＃階層クラスター

```
distance <- dist(DFfa)
#ユークリッド距離を求める

# 樹形図作成
hc <- hclust(distance, ”ward.D2”)
plot(hc)

res <- cutree(hc, k =4)
write.csv(res,”resultDF.csv”)
DF <- cbind(DF, res)　#列どうしを結合
```

＃k-means

```
kmDF <- kmeans(DFfa,4)
result_km <- kmDF$cluster
write.csv(result_km,”result_km.csv”)

#グラフ描画
library(cluster)
clusplot(DF, kmDF$cluster, color=TRUE,
shade=TRUE, labels=2, lines=0)
```

＃＃最終データセット

```
DF <- cbind(DF, result_km)　#列どうしを結合

write.csv(DF,”DF_all.csv”)
```

“DF_all.csv“を用いクラスター間で比較

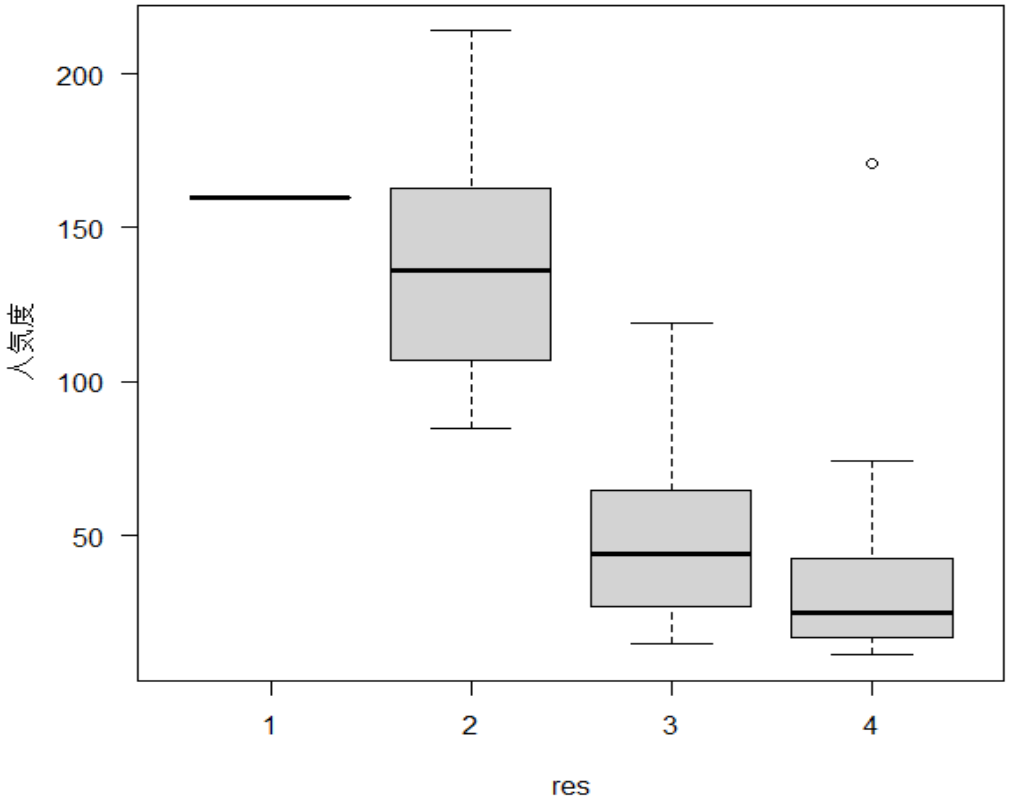
- EZRで実施

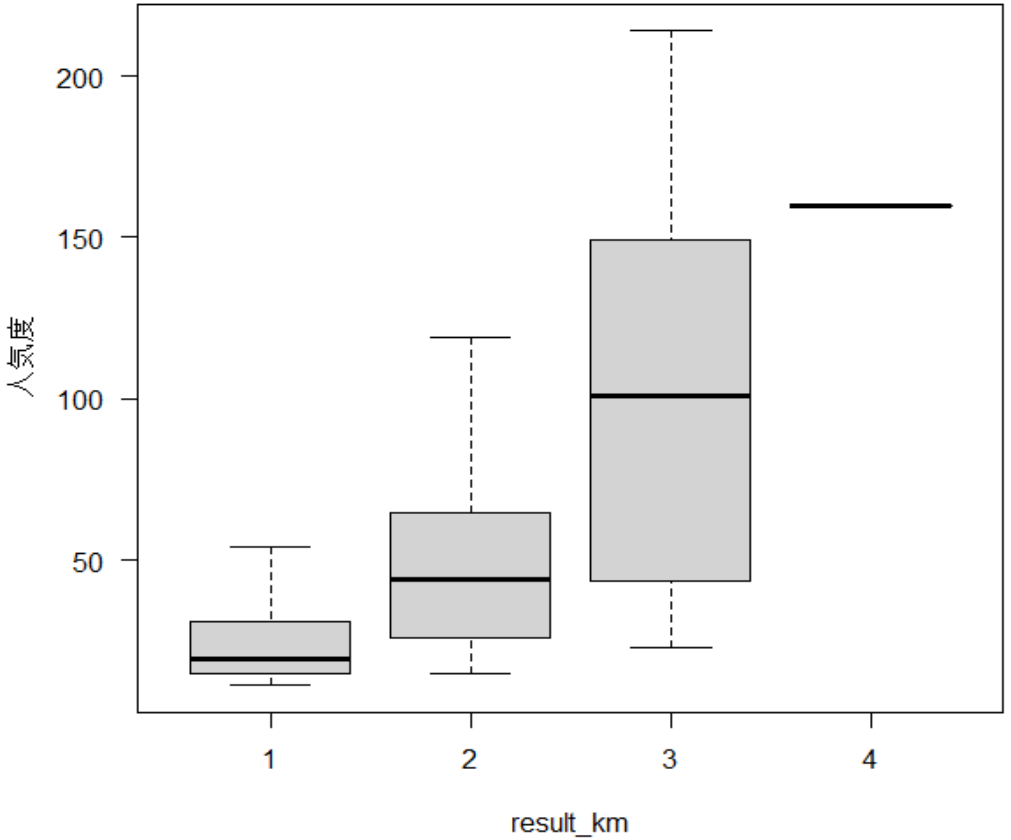
→[グラフ・表]-[背景データのサマリー表]

- Rcodeでそのまま実施でもよい

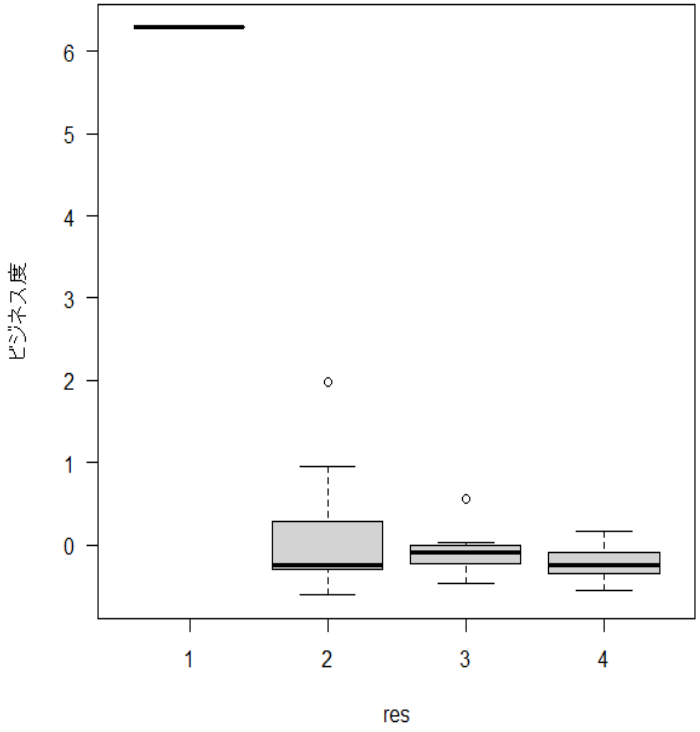
→library(tableone)

all <- CreateTableOne(vars = c(“人気度”, “ビジネス度”,“都会生活度”,“非高齢化度”),
strata=“res”,factorVars=c(“res”), data = DF)

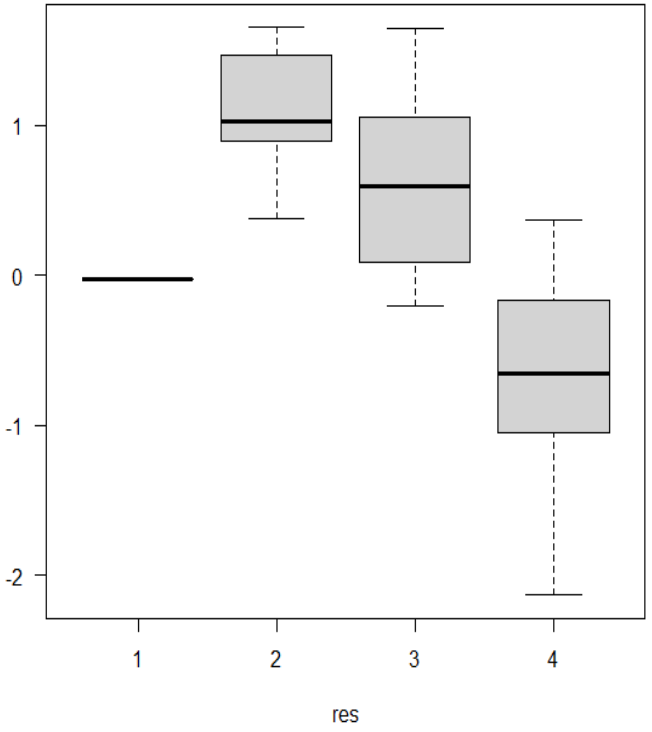

200
150
100
50
人気度
1
2
3
4
res


200
150
100
50
人気度
1
2
3
4
result_km

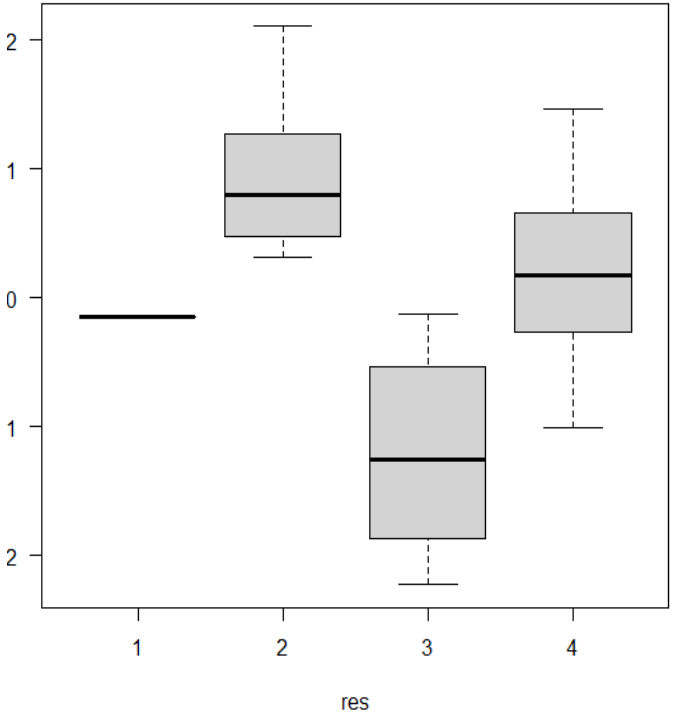
ビジネス度

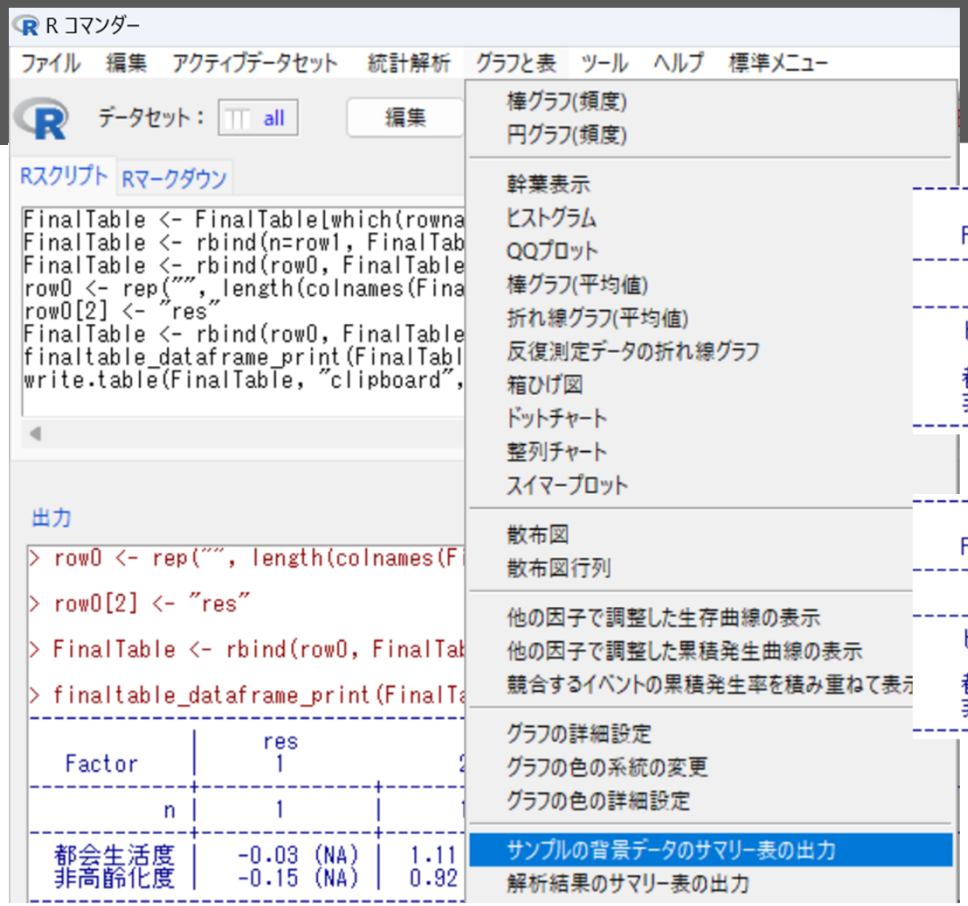


都会生活度



非高齢化度





Factor	res 1	2	3	4	p.value
n	1	10	12	27	
ビジネス度	6.29 (NA)	0.11 (0.80)	-0.10 (0.26)	-0.23 (0.20)	NA
人気度	160.00 (NA)	138.20 (37.96)	48.33 (29.30)	35.07 (32.25)	NA
都会生活度	-0.03 (NA)	1.11 (0.41)	0.63 (0.64)	-0.69 (0.69)	NA
非高齢化度	-0.15 (NA)	0.92 (0.58)	-1.22 (0.73)	0.21 (0.64)	NA

Factor	result_km 1	2	3	4	p.value
n	18	12	19	1	
ビジネス度	-0.14 (0.16)	-0.07 (0.24)	-0.15 (0.64)	6.29 (NA)	NA
人気度	24.39 (13.72)	48.17 (29.46)	99.58 (58.69)	160.00 (NA)	NA
都会生活度	-1.03 (0.58)	0.57 (0.68)	0.62 (0.62)	-0.03 (NA)	NA
非高齢化度	0.02 (0.56)	-1.27 (0.67)	0.79 (0.56)	-0.15 (NA)	NA

```
library(tableone)
all <- CreateTableOne(vars = c("人気度", "
ビジネス度","都会生活度","非高齢化度"),
strata="res",factorVars=c("res"),data = DF)
all
```

Stratified by res

	1	2	3	4	p	
n	1	10	12	27		
人気度 (mean (SD))	160.00 (NA)	138.20 (37.96)	48.33 (29.30)	35.07 (32.25)		NA
ビジネス度 (mean (SD))	6.29 (NA)	0.11 (0.80)	-0.10 (0.26)	-0.23 (0.20)		NA
都会生活度 (mean (SD))	-0.03 (NA)	1.11 (0.41)	0.63 (0.64)	-0.69 (0.69)		NA
非高齢化度 (mean (SD))	-0.15 (NA)	0.92 (0.58)	-1.22 (0.73)	0.21 (0.64)		NA

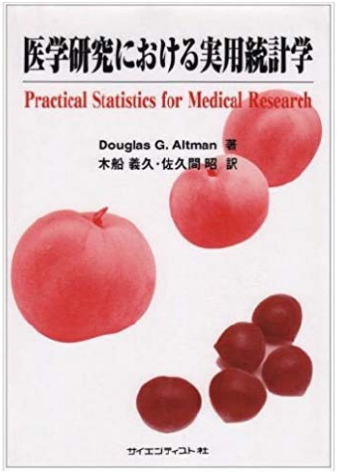
Stratified by result_km

	1	2	3	4	p	
n	18	12	19	1		
人気度 (mean (SD))	24.39 (13.72)	48.17 (29.46)	99.58 (58.69)	160.00 (NA)		NA
ビジネス度 (mean (SD))	-0.14 (0.16)	-0.07 (0.24)	-0.15 (0.64)	6.29 (NA)		NA
都会生活度 (mean (SD))	-1.03 (0.58)	0.57 (0.68)	0.62 (0.62)	-0.03 (NA)		NA
非高齢化度 (mean (SD))	0.02 (0.56)	-1.27 (0.67)	0.79 (0.56)	-0.15 (NA)		NA

第14回
2. 解析実習 Wrap up

教科書など

生物統計学(医学統計学)の学問体系



参考書
Altman DG: *Practical Statistics for Medical Research*,
Chapman and Hall,1991.
(木船義久、佐久間昭監訳:
「医学研究における実用統計学」、
サイエンティスト社、1999)

2

教科書(左)章立て例

データの型
データの記述
理論分布
研究計画(研究デザイン)
データ解析の準備
統計解析の原理
(推定、検定、モデル)
群間比較(連続データ)
群間比較(分類データ)
2連続変数間の関係
多変数間関係
生存時間解析
一致度、診断検査
臨床試験

大学学部の講義例

(1)バラツキとバイアス
(2)研究方法論
(3)評価の信頼性と妥当性
(4)検査データの解釈
(5)データの記述
(6)統計的推測
　　確率変数と確率分布
(7)　信頼区間とp値
(8)2群比較 2値データ
(9)2群比較 連続データ
(10)相関と回帰
　　相関の定義と解釈
(11) 回帰モデルの当てはめと診断
(12)多群の比較
(13)経時データ解析
(14)生存時間解析

おさえておいてほしい基礎

- 母集団と標本
- 正規分布と標準偏差
- 推定（点推定、区間推定）
- 統計学的検定
  - 帰無仮説
  - 検定の適用できるデータの特性
    - パラメトリック検定とノンパラメトリック検定
  - 2 群間の平均値の差の検定法（t 検定、Mann–Whitney U検定）
  - 多群の比較 （分散分析, Kruskal-wallis
  - 主な多重比較検定法（Dunnett 検定、Tukey 検定など）
- 解析方法
  - 相関
  - 重回帰分析: 最小二乗法による直線回帰
  - ロジスティック回帰
  - Poisson回帰
  - 生存時間解析（Kaplan–Meier 曲線, Cox回帰など）

教科書　（1） 統計学、生物統計学一般

（統計学一般）

- 東京大学教養学部統計学教室. 統計学入門. 東京大学出版会　1991
- 日本統計学会編.日本統計学会公式認定 統計検定2級対応　統計学基礎, 東京図書, 2015
- 江崎貴裕.分析者のためのデータ解釈学入門.ソシム,2020
- 阿部真人.データ分析に必須の知識・考え方　統計学入門. ソシム,2021

（生物統計学）

- Altman DG:Practical Statistics for Medical Research, Chapman andHall,1991.（佐久間昭監訳:「医学研究における実用統計学」、サイエンティスト社、1999）
- Armitage P and Berry G:Statistical Methods in Medical Research, 3rd ed. ,Blackwell, 1994.（椿美智子・椿広計共訳:「医学研究のための統計的方法」、サイエンティスト社、2001）
- 丹後俊郎:新版医学への統計学、朝倉書店、1993
- 浜田知久馬:学会・論文発表のための統計学新版、真興交易医書、2012
- 中村好一編.論文を正しく読み書くためのやさしい統計学 改訂第3版. 診断と治療社, 2019

教科書　（2）各種手法別

（統計パッケージ関連）

<SPSS>

- 対馬栄輝. 第2版　SPSSで学ぶ医療系データ解析.東京図書, 2016
- 対馬栄輝. 第2版　SPSSで学ぶ多変量医療系データ解析.東京図書, 2016

<SAS>

- 臨床評価研究会(ACE)基礎解析分科会 著.実用SAS　生物統計ハンドブック.サイエンティスト社
- 大橋渉.統計を知らない人のためのSAS入門 .オーム社, 2012

<R>

・笹渕 裕介, 大野 幸子, 橋本 洋平, 石丸 美穂. 超入門！すべての医療従事者のためのRstudioではじめる医療統計,金芳堂, 2021

<EZR>

- 神田善伸. EZRでやさしく学ぶ統計学 改訂3版 , 中外医学社、2020
- 新谷歩.みんなの医療統計 12日間で基礎理論とEZRを完全マスター! ,講談社, 2016

（疫学研究）

・スティーブン B. ハリー/スティーブン R. カミングス著. 木原雅子, 木原正博訳.医学的研究のデザイン　第4版.メディカルサイエンスインターナショナル, 2014

・Szklo Moyses, Nieto, F. Javier著アドバンスト分析疫学 369の図表で読み解く疫学的推論の論理と数理.メディカルサイエンスインターナショナル,2020