データマイニング

第４回　対応分析

2023年春学期

宮津和弘

本日の講義・演習

日付	講義・演習内容
04/14/23	(1) イントロダクション
04/21/23	(2) ビジネスシミュレーション
04/28/23	(3) ID-POSデータ分析
05/12/23	**(4) 対応分析**
05/19/23	(5) クラスター分析
05/26/23	(6) 自己組織化マップ
06/02/23	(7) 線形判別分析
06/09/23	(8) 非線形判別分析
06/16/23	(9) ツリーモデル
06/23/23	(10) 集団学習
06/30/23	(11) サポートベクターマシン
07/04/23	(12) ネットワーク分析
07/14/23	(13) 共分散構造分析
07/21/23	(14) テキスト分析
07/28/23	(15) まとめ

本日の演習概要とポイント

- 量的データ、質的データに対する分析手法
- サンプルデータの差の検定、データの次元削減
- 対応分析（コレスポンデンス分析、数量化Ⅲ類）

データの特性により統計手法は異なる

例）「TOEFL500点以上かどうか」と「TOEFL500点獲得した」では、データの意味は異なる！

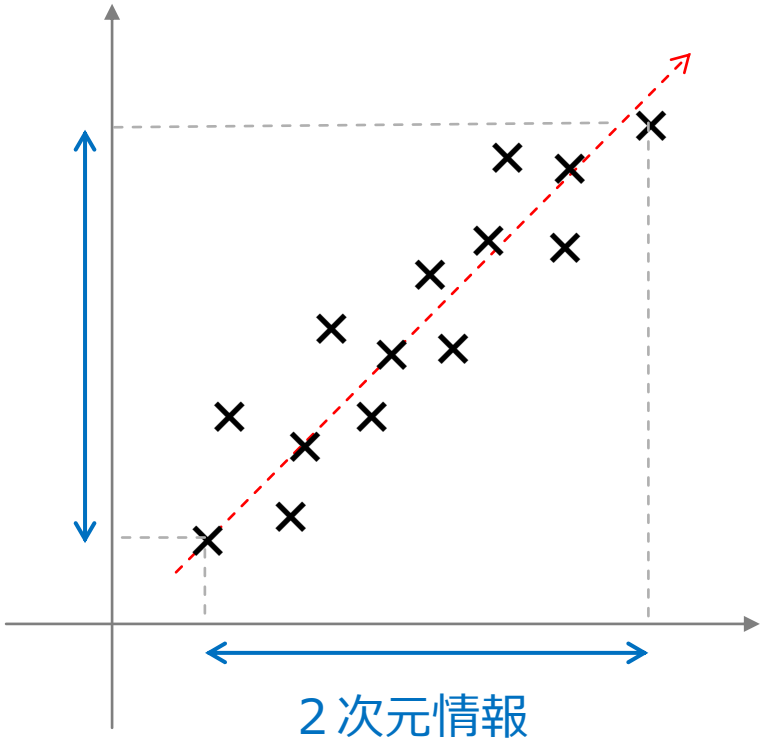
連続的データ(量的)で数値の大きさに意味がある
→　平均値、分散

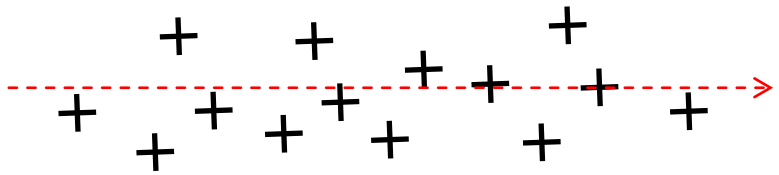
有無の意向(質的)を表し数値の大小に意味がない

	差の検定	次元削減
量的データ	t 検定	因子分析
質的データ	χ^2検定	対応分析

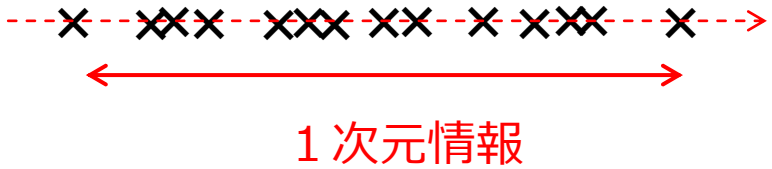
次元削減とは？

多次元情報を、元の意味を可能な限り維持しながら、より少ない次元情報で表す

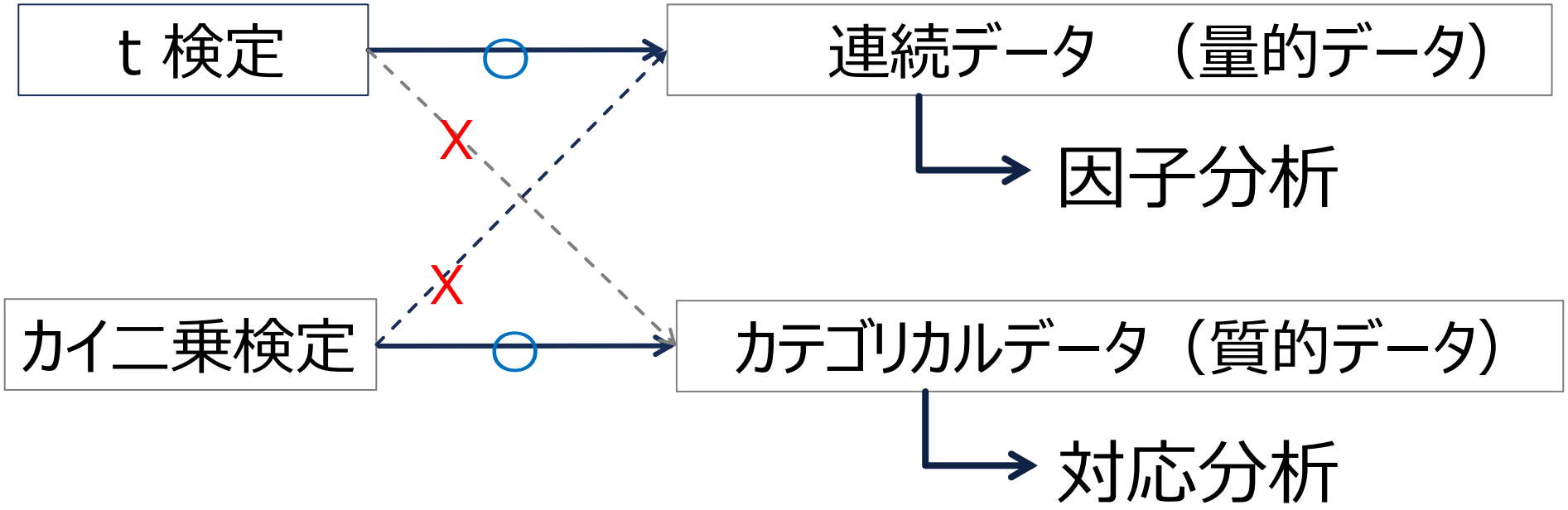




※　点線に沿った情報のばらつきが大きい



t 検定とカイ二乗検定の違い



分散分析（カイ二乗検定）

期待度数(E_{ij})に対して、実際の観測度数(n_{ij})が統計的有意に異なるかを検定する
⇒ 以下では、k × m の二元配置における独立性を考える

自由度$(k-1)(m-1)$のカイ二乗分布に従う

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2(k-1)(m-1)$$

	F_1	F_2	F_3	合計
R_1	n_{11}	n_{12}	n_{13}	r_1
R_2	n_{21}	n_{22}	n_{23}	r_2
合計	f_1	f_2	f_3	$k+m$

観測度数(n_{ij})

	F_1	F_2	F_3	合計
R_1	E_{11}	E_{12}	E_{13}	r_1
R_2	E_{21}	E_{22}	E_{23}	r_2
合計	f_1	f_2	f_3	$k+m$

期待度数(E_{ij})

カイ二乗分布

$N(\mu, \sigma^2)$ に従う確率変数 $x_1, x_2, x_3 \ldots, x_n$ に対して、正規化した変数の二乗和 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$はカイ二乗分布に従う

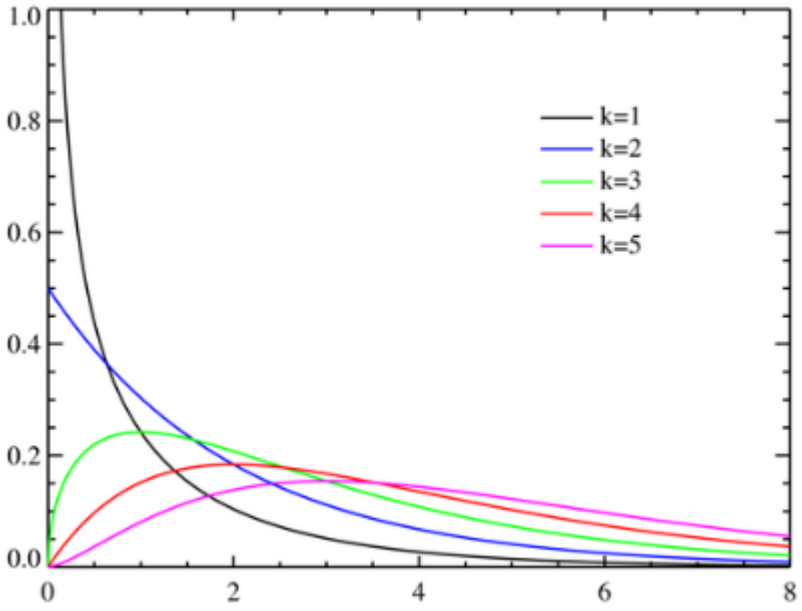
<u>カイ二乗分布の確率密度分布</u>

カイ二乗分布の確率密度関数は $x \geq 0$ に対し

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

また $x \leq 0$ に対し $f_k(x) = 0$ という形をとる。ここで Γ はガンマ関数である。

PDF of chi-square distribution



https://ja.wikipedia.org/wiki

どちらのデジカメが好みですか？

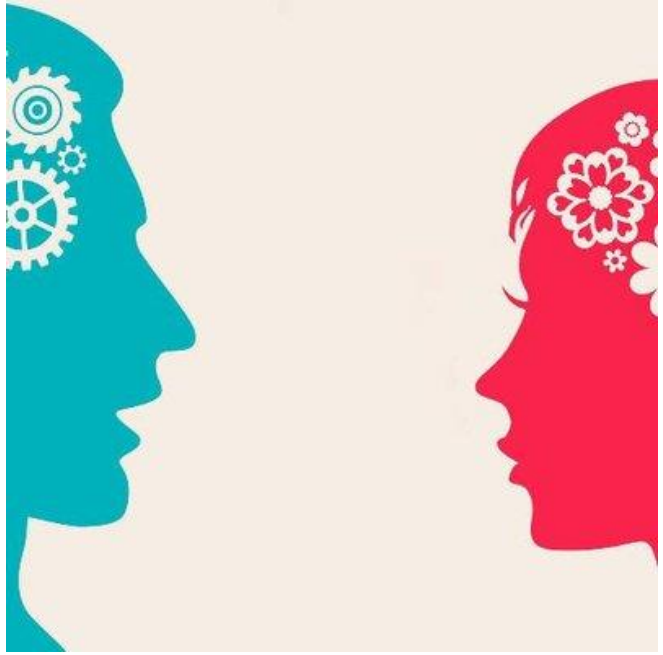
どちらのデジカメが好みか、大学生の男女100人に聞きました。

vs.

クロス集計表：デジカメの好みに男女差はあるか？

以下の結果から、２つのデジカメの好みに男女差があると言えるか？

	男性	女性	合計
	30	22	52
	23	25	48
合計	53	47	100



観測値と期待値

観測値

	男性	女性	合計
	30	22	52
	23	25	48
合計	53	47	100

男性と女性でデジカメの好みに差異が無ければ、男女とも同じ割合で観測されるはずである。

期待値

	男性	女性	合計
	27.56	24.44	52
	25.44	22.56	48
合計	53	47	100

$$52 \times 0.53 = 27.56 \qquad 52 \times 0.47 = 24.44$$

$$48 \times 0.53 = 25.44 \qquad 48 \times 0.47 = 22.56$$

カイ二乗検定

```
> A<-c(30,22)
> B<-c(23,25)
> camera<-rbind(A,B)
> camera
  [,1] [,2]
A   30   22
B   23   25
```

	男性	女性	合計
	30	22	52
	23	25	48
合計	53	47	100

```
> chisq.test(camera)

        Pearson's Chi-squared test with Yates' continuity correction

data:  camera
X-squared = 0.60532, df = 1, p-value = 0.4366
```

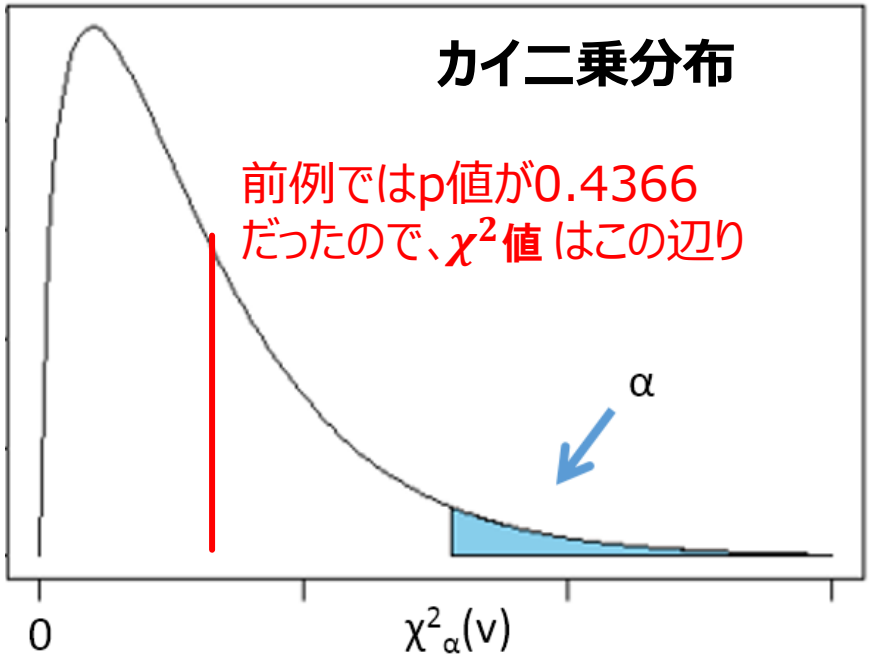
95%有意性で異ならない → **デジカメの好みに差がない！**

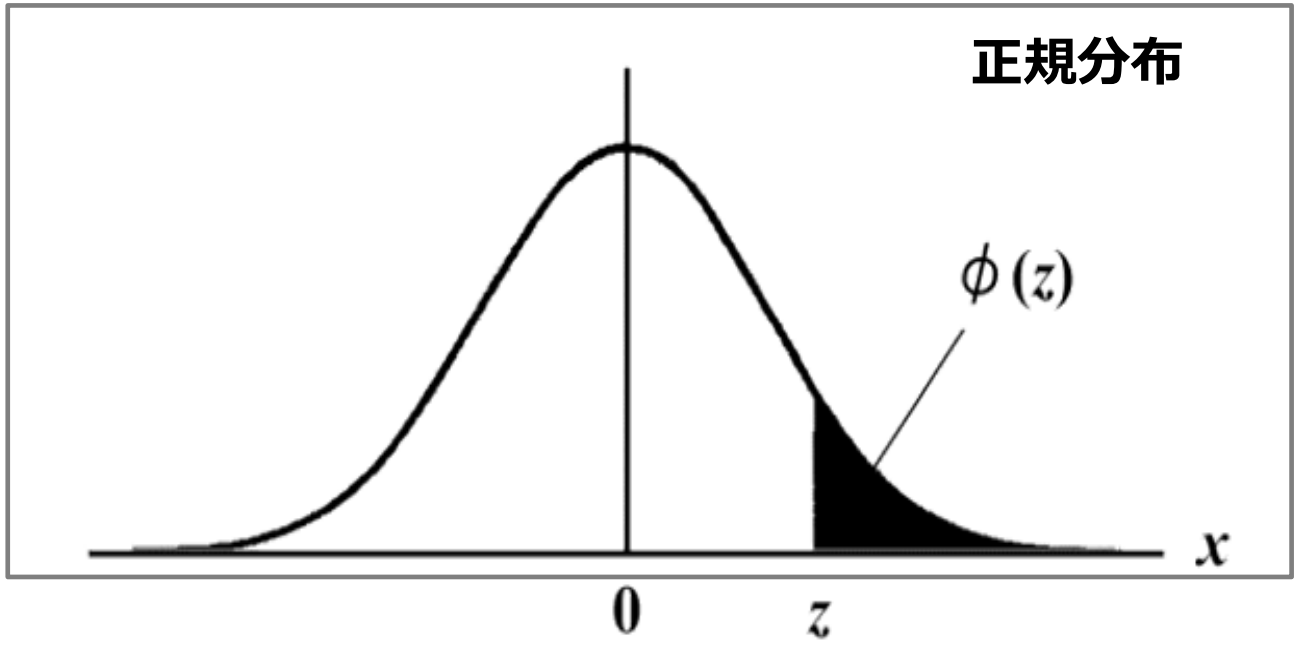
イェーツ連続補正：データが少ない場合等で検定を厳しくする

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(|n_{ij}-E_{ij}|-0.5)^2}{E_{ij}}\sim\chi^2(k-1)(m-1)$$

統計的有意性の解釈

確率密度分布は、-∞～+∞で積分すると１となる関数である。以下の確率密度関数の網掛け部分が0.05のとき95%の確率的有意性で起こりえないこと、白い部分については95%で起こり得ることと捉える。この網掛け部分の数値がp値で、これが0.05より大きいと、確率的有意性95%で起こり得る（前例では、異ならないということ！）





質的データと量的データの分析手法

大学生男女100人に対して、デジカメ商品AとBの評価を５尺度で点数をつけてもらった。

量的データ

男性	3.85	3.06
女性	3.10	3.27

商品と性別ごとに平均値を算出する

⇒ **因子分析**

回答者	商品A	商品B
1	3	5
2	2	3
3	5	4
4	1	4
5	2	1
6	4	3
7	3	5
8	4	1
9	2	2
10	5	3
11	1	2
12	3	3
⋮	⋮	⋮
N	3	5

質的データ

男性	30	23
女性	22	25

商品と性別ごとに４以上評価をカウントする

⇒ **対応分析**

対応分析とは？

カイ二乗検定が質的データに対する差異の有無を検定するように、対応分析も質的データに対して行う分析手法の一つである。質的データは量的データのように値の大きさに意味を持たない。しかし、**対応分析** では**質的データに潜在的な量的データを割り当てる**ことで量的なデータ処理が可能となる。

例）

	男性	女性	合計
	30	22	52
	23	25	48
合計	53	47	100

← 大学生の男女100人に対して、
どちらのデジカメが好みかのアンケート

対応分析のデータ収集例

アンケート調査(N人対象)を実施して、年代ごとに好みのシャンプーブランドを選択する

アンケート結果

回答者	年齢	ブランド
1	50以上	T
2	34以下	L
3	34-49	M
4	34以下	L
5	50以上	T
⋮	⋮	⋮
419	34-49	M

クロス集計結果

	L	M	P	T
34以下	27	19	38	26
34-49	46	31	34	65
50以上	43	41	20	29

⇒ **質的データでは定量分析ができない！**

対応分析における潜在変数の導入

質的データ(年齢、ブランド)が定量評価できるように、背後に潜在定量データを仮定する

年齢	潜在変数
34以下	$v1$
39-49	$v2$
50以上	$v3$

ブランド	潜在変数
L	$w1$
M	$w2$
P	$w3$
T	$w4$

この仮定より、以下のように定量化が可能となる

$$M_x = \frac{n_1 v_1 + n_2 v_2 + n_3 v_3}{N}$$

$$M_y = \frac{n_1 w_1 + n_2 w_2 + n_3 w_3}{N}$$

$$s_x^2 = \frac{n_1 v_1^2 + n_2 v_2^2 + n_3 v_3^2}{N} - M_x^2$$

$$s_y^2 = \frac{n_1 w_1^2 + n_2 w_2^2 + n_3 w_3^2}{N} - M_y^2$$

質的データ数量化の仮定

クロス集計結果

	L	M	P	T
34以下	n_{11}	n_{12}	n_{13}	n_{14}
35-49	n_{21}	n_{22}	n_{23}	n_{24}
50以上	n_{31}	n_{32}	n_{13}	n_{34}

（総サンプル数） $N = \sum_{j=1}^{4} \sum_{i=1}^{3} n_{i,j}$

+

回答者	年齢 x	ブランド y
1	$v3$	$w4$
2	$v1$	$w1$
3	$v2$	$w2$
4	$v1$	$w1$
5	$v3$	$w4$
⋮	⋮	⋮
N	$v2$	$W2$

$$M_x = M_x = 0$$

$$s_x^2 = s_x^2 = 1$$

x, y の平均 0 および分散 1

対応分析の定式化

以下の条件の下、

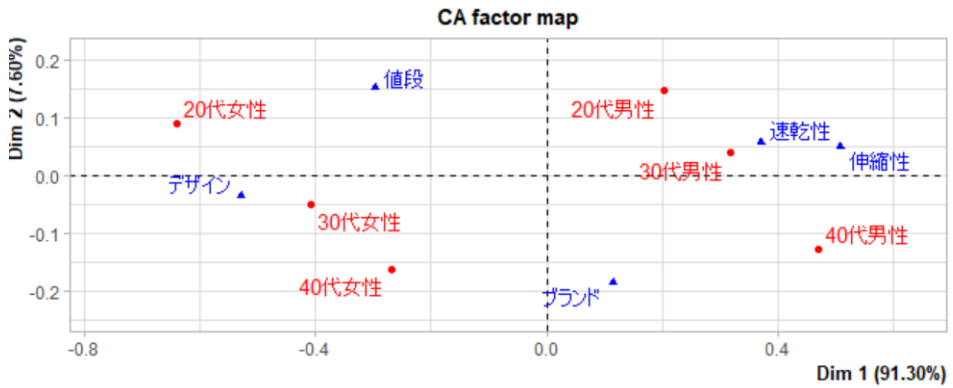
$$n_{1.}v_1^2 + n_{2.}v_2^2 + \cdots + n_{l.}v_l^2 = N$$
$$n_{.1}w_1^2 + n_{.2}w_2^2 + \cdots + n_{.l}w_m^2 = N$$

$\boldsymbol{r}_{xy}$ を最大にする $\boldsymbol{v}, \boldsymbol{w}$ を求める

$$r_{xy} = n_{11}v_1w_1 + \cdots + n_{1m}v_1w_m + \cdots + n_{l1}v_lw_1 + \cdots n_{lm}v_lw_m$$

$\boldsymbol{H} = \boldsymbol{G}^T\boldsymbol{G}$の固有値と固有ベクトルを$\boldsymbol{\lambda}$および$\boldsymbol{z}$のとき

$$G = \begin{pmatrix} \frac{n_{11}}{\sqrt{n_{1.}n_{.1}}} & \cdots & \frac{n_{1m}}{\sqrt{n_{1.}n_{.m}}} \\ \vdots & \ddots & \vdots \\ \frac{n_{l1}}{\sqrt{n_{l.}n_{.1}}} & \cdots & \frac{n_{lm}}{\sqrt{n_{l.}n_{.m}}} \end{pmatrix}$$



$$w_j = \frac{z_j}{\sqrt{n_{.j}}} \quad (j = 1,2,\ldots,\mathrm{l})$$

$$v_i = \frac{n_{i1}w_1 + \cdots + n_{im}w_m}{n_{i.}\sqrt{\lambda}} \quad (i = 1,2,\ldots,\mathrm{l})$$

FactoMineR のインストール

※　CRANのミラーサイトはJapanを選択

```
install.packages("Rcmdr")
install.packages("FactoMineR")
install.packages("RcmdrPlugin.FactoMineR")
```

Rのコンソール画面に下記のように入力します。

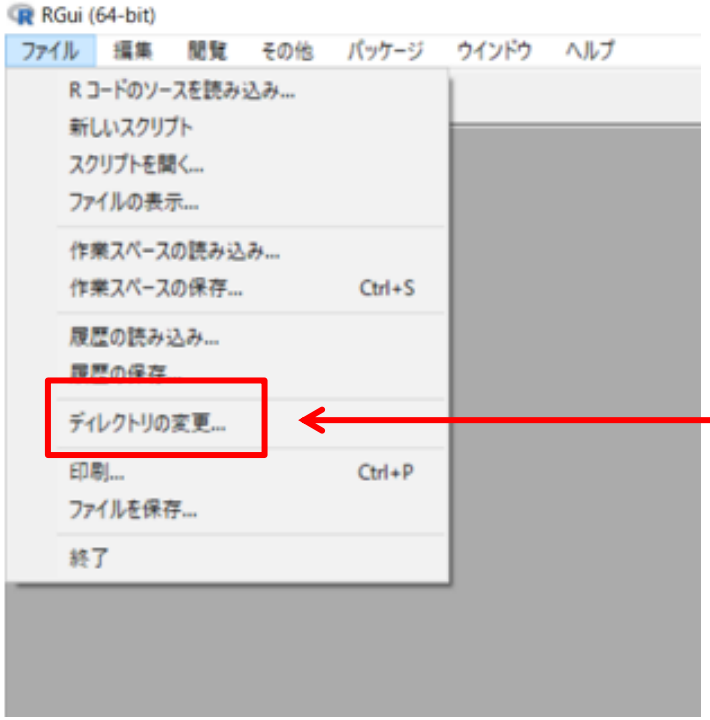
```
library(Rcmdr)
```

Rコマンダーの画面が表示されます。
（ツール）のRcmdrプラグインのロードを選んで、
RcmdrPlugin.FactoMineRを選択します。

再起動後、Rコマンダーのメニューに**FactoMineR**が表示されます。

※ Rcmdr起動時、毎回ツールからFactoMineRを立ち上げる

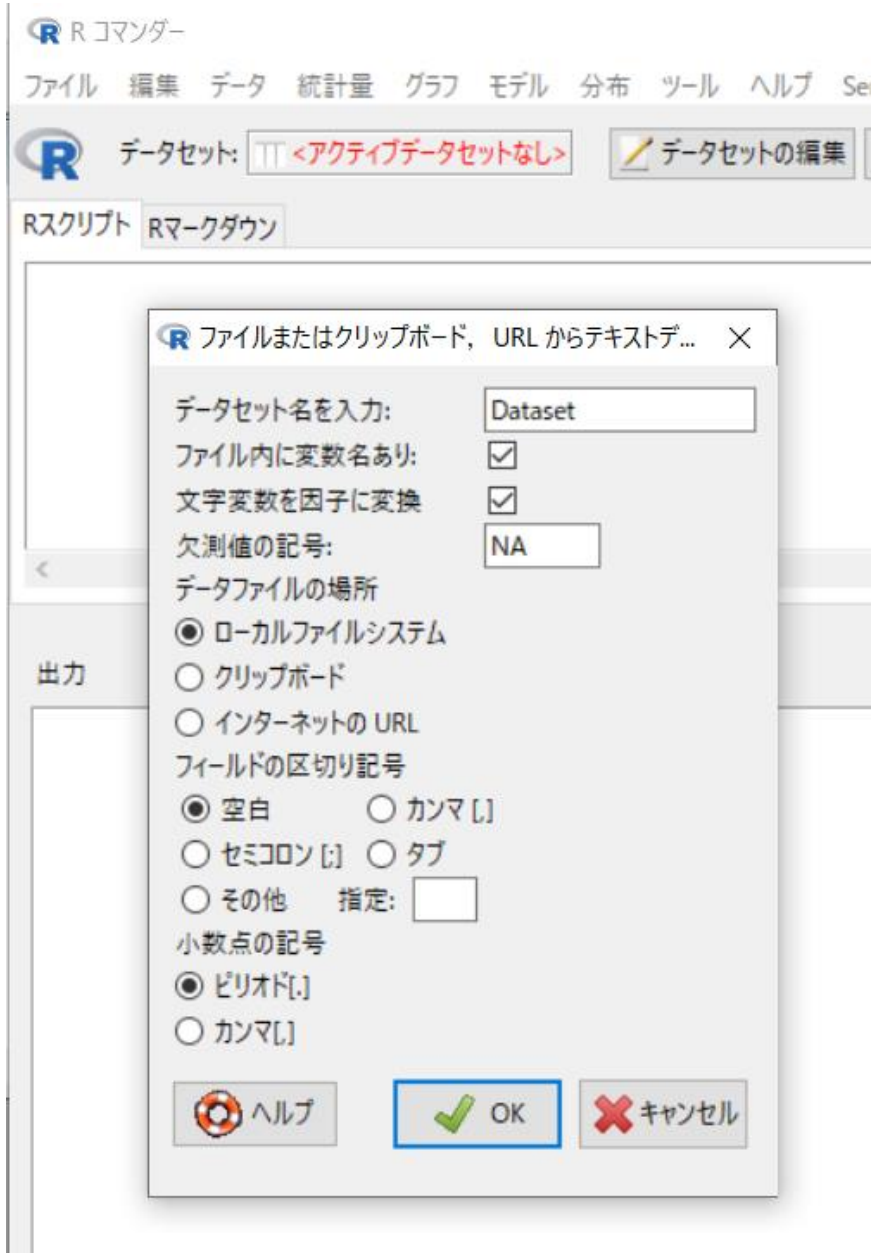
ファイルの入出力を指定する



演習データも保存しておく！

- **Camera.txt**

R_workingなどを作成して、ワーキングディレクトリとして指定する



Rコマンダー

データ ⇒ データのインポート
⇒ テキストファイルまたはクリップボード
⇒ OK

作業ディレクトリから演習ファイルを選択

- Camera.txt

データセットがロードされる

演習データの確認

⇒ データセットの編集/表示

10種類のカメラ製品に対する評価データ
500人が各カメラの項目について〇または×をつけた
以下の表は集計データを100で除した値である

データエディタ: Dataset

ファイル　編集　ヘルプ

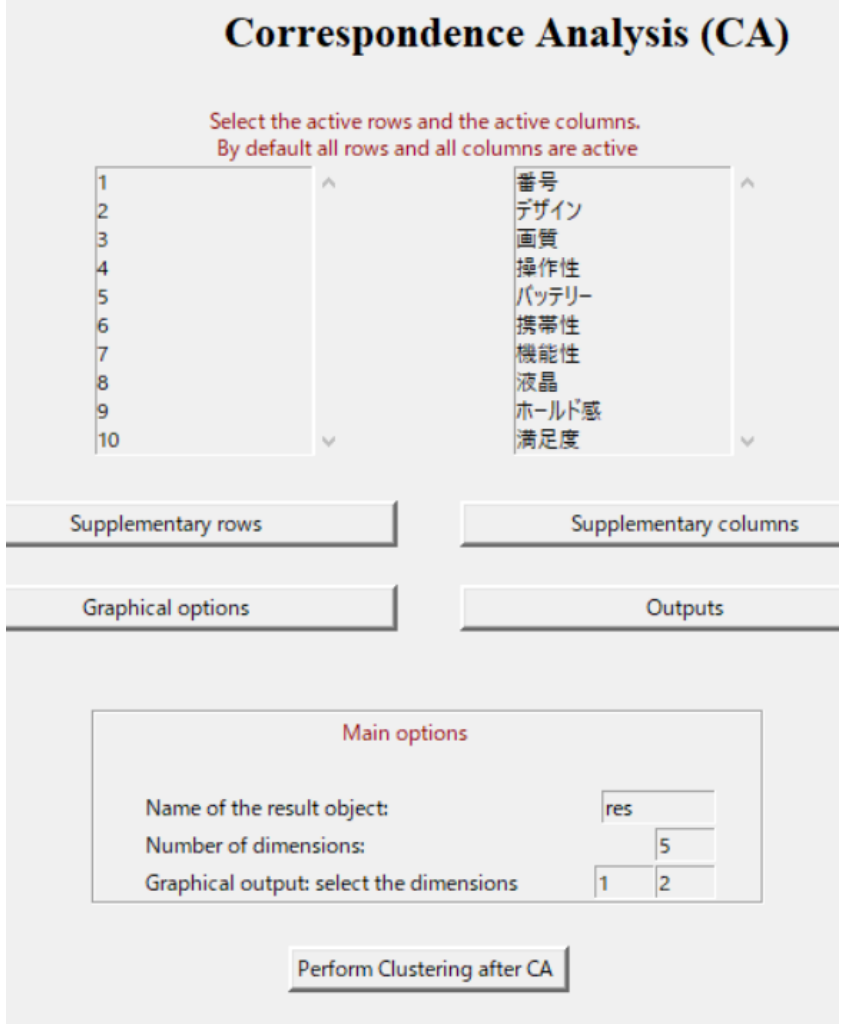
行の追加　列の追加

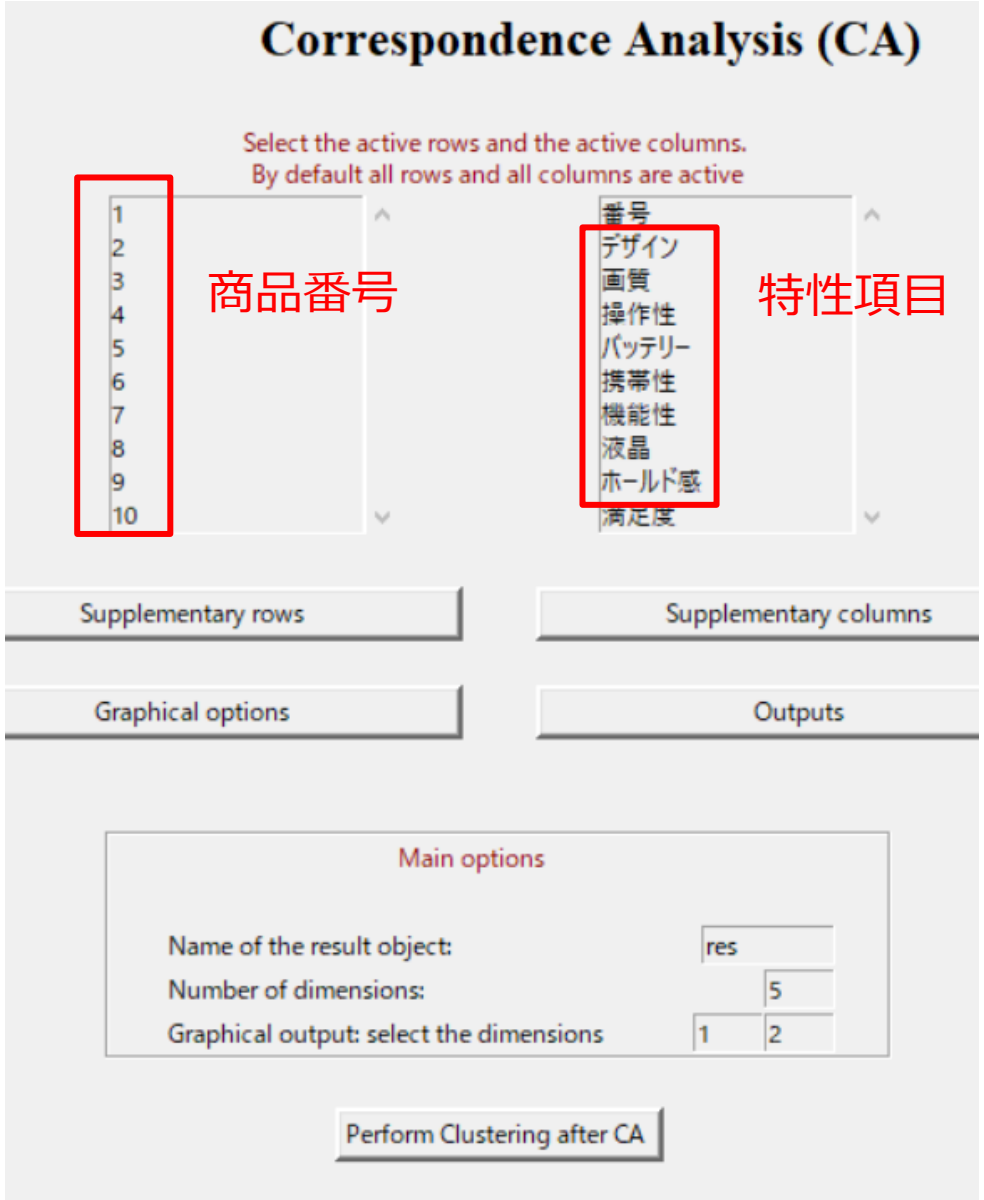
		1	2	3	4	5	6	7	8	9	10
	rowname	番号	デザイン	画質	操作性	バッテリー	携帯性	機能性	液晶	ホールド感	満足度
1	1	1	4.78	4.77	4.72	4.67	4.64	4.72	4.84	4.84	4.97
2	2	2	4.8	4.8	4.4	4	5	4.8	4.9	4.4	4.9
3	3	3	4.02	3.95	3.83	2.71	3.3	4.03	4.71	3.23	4.64
4	4	4	4.72	4.66	4.38	4.29	4.14	4.77	4.54	4.39	4.7
5	5	5	4.6	4.66	4.24	3.83	4.42	4.63	4.72	4.79	4.81
6	6	6	4.54	4.77	3.69	4.22	4.22	4.51	4.28	3.96	4.76
7	7	7	4.51	4.43	4.13	4.18	4.77	4.45	4.38	3.85	4.72
8	8	8	4	4.9	4.3	2.05	4.11	3.51	4.1	3.61	4.77
9	9	9	4.73	4.45	4.64	4.64	3.91	4.64	4.73	5	4.73
10	10	10	4.64	4.82	4.11	4.36	4.67	4.72	4.91	3.48	4.72

ヘルプ　OK　キャンセル

CA（対応分析）起動

FactoMineRから
Correspondence Analysis (CA) を選択





各商品と特性項目を同じ
平面上で同時に評価する

行の分析結果

二次元変換後、Dim.1とDim.2で元の78%を表せる

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Variance	0.004	0.001	0.001	0.000	0.000	0.000	0.000
% of var.	55.391	22.440	14.205	4.045	3.222	0.611	0.085
Cumulative % of var.	55.391	77.831	92.036	96.082	99.304	99.915	100.000

Rows

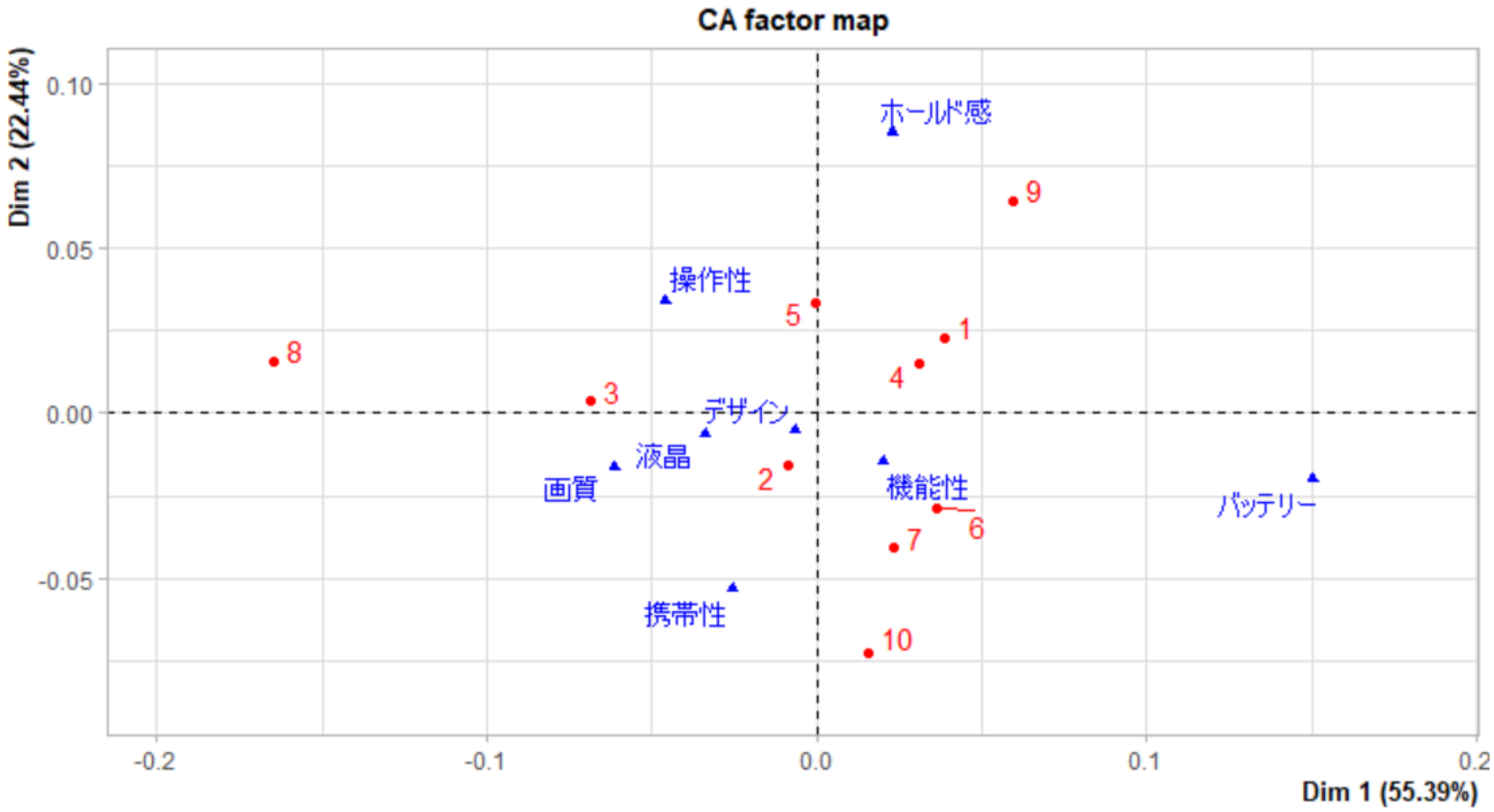
	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
1	0.282	0.039	4.511	0.581	0.023	3.905	0.204	0.015	2.493	0.082
2	0.135	-0.008	0.210	0.056	-0.016	1.857	0.202	0.015	2.586	0.178
3	1.042	-0.068	10.985	0.383	0.004	0.075	0.001	-0.086	68.303	0.611
4	0.174	0.031	2.786	0.581	0.015	1.601	0.135	-0.006	0.433	0.023
5	0.220	0.000	0.001	0.000	0.033	7.844	0.525	0.012	1.492	0.063
6	0.371	0.036	3.568	0.349	-0.029	5.479	0.217	0.011	1.305	0.033
7	0.325	0.023	1.491	0.167	-0.041	11.347	0.514	0.023	5.851	0.168
8	2.542	-0.165	65.434	0.935	0.016	1.513	0.009	0.037	12.593	0.046
9	0.849	0.060	10.324	0.442	0.065	29.854	0.518	-0.011	1.328	0.015
10	0.618	0.016	0.691	0.041	-0.072	36.526	0.870	-0.018	3.617	0.055

列の分析結果

※　行と列で同じ軸で表している

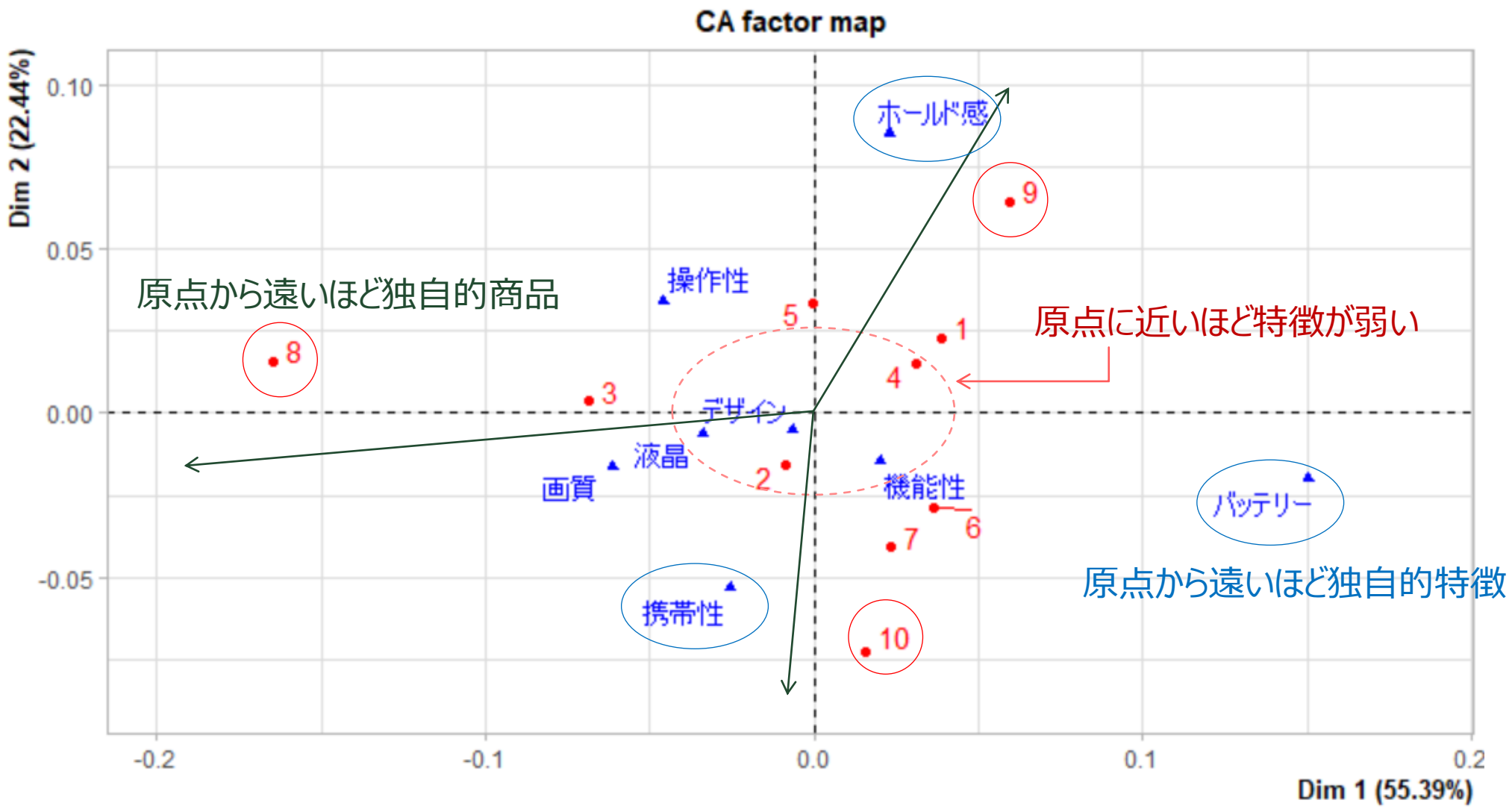
Columns

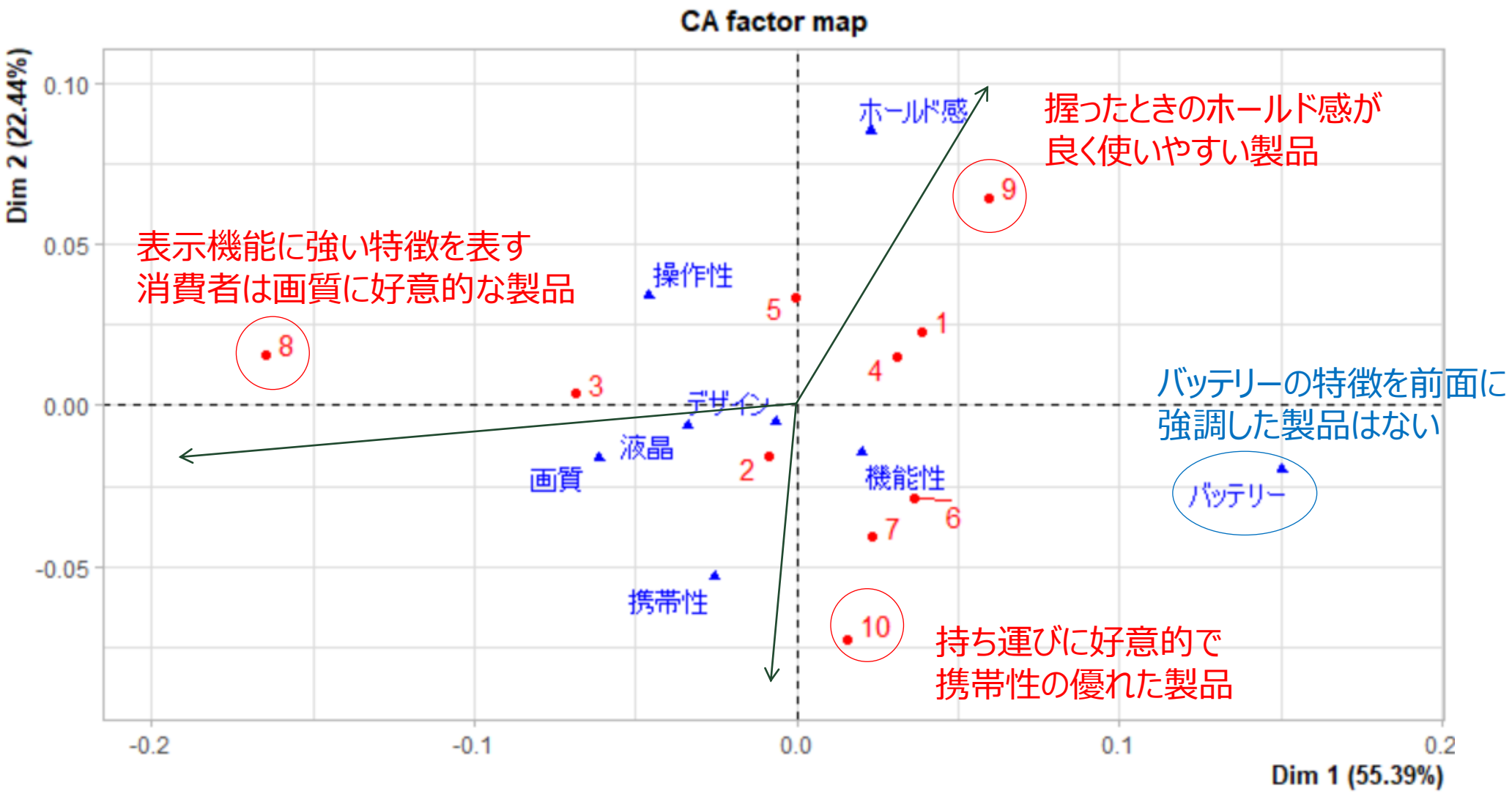
	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
デザイン	0.043	-0.006	0.151	0.127	-0.005	0.248	0.085	-0.010	1.448	0.313
画質	0.729	-0.061	13.724	0.684	-0.016	2.359	0.048	0.019	4.928	0.063
操作性	0.530	-0.046	7.010	0.480	0.034	9.527	0.264	-0.003	0.092	0.002
バッテリー	2.600	0.150	69.606	0.972	-0.020	3.021	0.017	0.002	0.054	0.000
携帯性	0.809	-0.025	2.184	0.098	-0.053	23.676	0.431	0.050	32.784	0.378
機能性	0.215	0.020	1.462	0.247	-0.014	1.832	0.126	-0.026	9.530	0.413
液晶	0.577	-0.034	4.152	0.262	-0.006	0.338	0.009	-0.054	41.323	0.668
ホールド感	1.055	0.023	1.712	0.059	0.085	59.000	0.823	0.028	9.840	0.087



※　製品番号(●)と商品特性(▲)を同一軸を用いて、同一面で評価

商品と特性を
同一平面に
プロットする


CA factor map
Dim 2 (22.44%)
Dim 1 (55.39%)
0.10
0.05
0.00
-0.05
-0.2
-0.1
0.0
0.1
0.2
ホールド感
9
操作性
原点から遠いほど独自的商品
5
1
原点に近いほど特徴が弱い
8
4
3
デザイン
液晶
画質
2
機能性
バッテリー
6
7
原点から遠いほど独自的特徴
携帯性
10


CA factor map
Dim 2 (22.44%)
Dim 1 (55.39%)
0.10
0.05
0.00
-0.05
-0.2
-0.1
0.0
0.1
0.2
ホールド感
操作性
デザイン
液晶
画質
機能性
携帯性
バッテリー
1
2
3
4
5
6
7
8
9
10
握ったときのホールド感が
良く使いやすい製品
表示機能に強い特徴を表す
消費者は画質に好意的な製品
バッテリーの特徴を前面に
強調した製品はない
持ち運びに好意的で
携帯性の優れた製品

課題：スポーツウエアのコレスポンデンス分析

20～40代男女各100名に対して、スポーツウエアを購入する際のポイントをアンケート調査した。
あてはまるものをすべて（複数回答）選んでもらった結果が以下の表である。この結果に対してコレスポンデンス分析を実施して、結果を考察せよ。

	伸縮性	デザイン	速乾性	ブランド	値段
20代男性	68	52	82	56	62
20代女性	11	89	13	29	61
30代男性	65	38	73	60	41
30代女性	13	74	25	41	46
40代男性	67	24	62	71	23
40代女性	18	72	31	53	38

スポーツウエア.csv

データマイニングを楽しもう！